

Computer Engineering
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ce.et.tudelft.nl/



CE-2012-17

M.Sc. Thesis

Power efficient digital correlator in the scope of an UWB baseband design

Jing Cao B.Sc.

Abstract

Ultra Wideband (UWB) radio represents a promising way of communication for low power applications in interference-prone environments. For achieving a low power solution the digital baseband architecture needs to be carefully optimized to reduce the total power consumption. Particularly for the computationally intensive synchronization phase that detects UWB signals in noisy input data, low-power correlation architectures are crucial. In this thesis, a low power exploration is performed at different levels of abstraction. The architecture of the correlator is designed and optimized to support several modes of operations efficiently. Deviating from a memory dominant design, our improved architecture is based on circular register buffer and a corresponding partitioning method. The impact of biased representation is studied in terms of power and area. The possibility and the influence of voltage reduction is also investigated by changing to the Lvt cells and pipelining the critical path. Our design is implemented and simulated in a Cadence based design flow, targetting 90 nm process technology. Experimental results show that, compared to "standard" design, the proposed correlator can achieve up to 2x total power reduction with only 10% overhead on area.

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science Delft University of Technology

Power efficient digital correlator in the scope of an UWB baseband design

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Embedded System

by

Jing Cao B.Sc.
born in Nanjing, China

This work was performed in:

Computer Engineering Group
Department of Microelectronics & Computer Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Copyright © 2012 Computer Engineering Group
All rights reserved.

Delft University of Technology
Department of
Microelectronics & Computer Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Power efficient digital correlator in the scope of an UWB baseband design"** by **Jing Cao B.Sc.** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: Oct 24th 2012

Chairman:

prof.dr. Koen Bertels

Advisors:

ir. Benjamin Büsze

dr.ir. Georgi Gaydadjiev

Committee Members:

prof.dr. Paddy French

dr. Maryam Ashouei

Abstract

Ultra Wideband (UWB) radio represents a promising way of communication for low power applications in interference-prone environments. For achieving a low power solution the digital baseband architecture needs to be carefully optimized to reduce the total power consumption. Particularly for the computationally intensive synchronization phase that detects UWB signals in noisy input data, low-power correlation architectures are crucial. In this thesis, a low power exploration is performed at different levels of abstraction. The architecture of the correlator is designed and optimized to support several modes of operations efficiently. Deviating from a memory dominant design, our improved architecture is based on circular register buffer and a corresponding partitioning method. The impact of biased representation is studied in terms of power and area. The possibility and the influence of voltage reduction is also investigated by changing to the Lvt cells and pipelining the critical path. Our design is implemented and simulated in a Cadence based design flow, targetting 90 nm process technology. Experimental results show that, compared to "standard" design, the proposed correlator can achieve up to 2x total power reduction with only 10% overhead on area.

Acknowledgments

First and foremost, I am extremely grateful to my guide, Benjamin Büsze, for his support and suggestions during the eight-month of thesis research. He gives me such a good opportunity to do my master thesis work at a stimulating and learning environment, imec-NL/Holst Centre. Without his patient guidance and constant encouragement, the thesis work would not have been done. His knowledge and passion for working set a good example for me in future career.

I have been privileged to interact with the other colleagues, Dr. Chris Christian Bachmann, Dr. Maryam Ashouei and Alex Young, for their insightful reviews and suggestions during the thesis.

Special thanks are given to my advisor in the university, Dr. Georgi Gaydadjiev, for his kindness and valuable feedback during the meetings and thesis review. I would like to appreciate Prof. Koen Bertels and Prof. Paddy French can share his time and become one of the committee member.

I want to thank other Chinese friends, Ao Ba, Ming Ding, Di Li, I appreciate their help and companionship during the entire thesis duration.

Last but not least, I would also like to express my thanks to my parents and grandparents. They always give me endless love and support, especially when I was once dropped out the university for severe back problem. Their encouragement is exactly the engine and power of entire master study.

Jing Cao B.Sc.
Delft, The Netherlands
Oct 24th 2012

Contents

Abstract **v**

Acknowledgments **vii**

1 Introduction **1**
1.1 Motivation . . . 2
1.2 Contribution . . . 3
1.3 Outline of the work . . . 4

2 Background **5**
2.1 UWB principle . . . 5
2.1.1 UWB definition . . . 5
2.1.2 Impulse Radio . . . 6
2.1.3 Frame Structure . . . 7
2.2 UWB Synchronization . . . 9
2.2.1 Synchronization procedure . . . 9
2.2.2 Types of preamble detection . . . 9
2.2.3 Principle of correlation . . . 10
2.3 ASIC design flow . . . 11
2.3.1 RTL behavioral description and simulation . . . 11
2.3.2 Logic synthesis and simulation . . . 12
2.3.3 Physical design. . . . 12
2.4 Overview of low power design . . . 13
2.4.1 Low power design methodology . . . 14
2.4.2 Power analysis procedure . . . 15
2.5 Literature research . . . 16

3 Low power Exploration **19**
3.1 Preamble detection procedure . . . 20
3.2 Architecture optimization . . . 21
3.2.1 Architecture comparison . . . 22
3.2.2 Overview of the optimized architecture . . . 24
3.2.3 Data path optimization in *delcorr* module . . . 25
3.3 Design of sub-sequence delay line . . . 26
3.3.1 Access pattern changing on SRAM . . . 27
3.3.2 Comparison on Flip-flops and SRAM based circular buffer . . . 28
3.3.3 Register partitioning . . . 28
3.4 Arithmetic Level exploration . . . 30
3.5 Circuit Level exploration . . . 31

4 Evaluation and Experimental Results **33**
4.1 Evaluation on power optimizations 33
4.1.1 Access pattern change 33
4.1.2 SRAM and register based circular buffer 34
4.1.3 Architecture optimization 34
4.1.4 Register partitioning 36
4.1.5 Biased representation on sub-channels 37
4.1.6 Voltage scaling 38
4.2 Final power summary 40
4.3 Physical layout view 42

5 Conclusion and Future work **45**
5.1 Conclusion 45
5.2 Future Work 45

Bibliography **47**

List of Figures

1.1 Applications on UWB Technology. 1
1.2 Comparison between typical wireless standards [5] 2
1.3 Power profiling of an UWB frame reception(source: Holst Centre/imec) 3

2.1 Ultra-wideband frequency principle 5
2.2 Ultra-wideband FCC spectral emission limits [8] 6
2.3 Duty cycle difference 7
2.4 UWB PHY frame structure. 7
2.5 SHR symbols. 8
2.6 Short preamble code sequences (source: Holst Centre/imec). 8
2.7 Overview of receiver in UWB system. 9
2.8 Spreading code phase detection. 11
2.9 Structure of digital matched filter. 11
2.10 ASIC design flow. 12
2.11 Dynamic power & Static power 13
2.12 Power tradeoff between different low power techniques 15
2.13 Power Analysis Flows [17] 16

3.1 Levels of design abstraction for power saving [25] 19
3.2 Preamble detection flow 20
3.3 Position determination process of the accumulator(source: Holst Centre/imec) 20
3.4 Direct form of preamble correlator 21
3.5 Two kinds of FIR achitecutre 22
3.6 Segmented form of FIR filter with serial delay lines 23
3.7 Overview of correlator architecture 24
3.8 Detailed data path for summation and magnitude calculation 26
3.9 Memory access pattern based on two pointers 27
3.10 Memory access pattern based on one pointer 27
3.11 Integrated clock gating circuit [28] 28
3.12 Memory usage in sub-delay line 29
3.13 Partitioned registers with clock gating 29
3.14 Partitioned registers with clock gating and operand isolation 30
3.15 Critical path in the correlator under 0.7V cell library 32

4.1 Power comparison between SRAM and DFFs 34
4.2 Power results on different FIR architectures 35
4.3 Power distribution in direct and segmented FIR architectures 35
4.4 Power results on partitioned delay line 36
4.5 Power and area changes of sub-channels in different optimization stages 37
4.6 Power reduction by inserting pipeline 38
4.7 Power analysis after the reduced power supply 39
4.8 Normalized area change during the voltage exploration 39

4.9 Power results in different stages 41
4.10 Power distribution comparison 42
4.11 Physical Layout of the design 43

List of Tables

3.1 Representation of ternary multiplication in *subcorr* module 25

4.1 Power comparison between the two access patterns 33
4.2 Comparison between parallel and serial delay line 36
4.3 Final power results summary 41

1 Introduction

Ultra Wideband (UWB) radio is one of the emerging communication technologies for wireless personal area networks. In contrast to traditional narrow and wideband systems, UWB uses a much wider frequency band. Where traditional radio systems use bandwidths from several kHz to a few MHz, UWB uses a much larger bandwidth (i.e. larger than 500 MHz) so that it can better handle large capacities of information. Besides, it can also provide attractive features, multipath immunity, and low power spectral density [1]. These distinct characteristic together with proliferation of wireless communication motivate more interest in practical use of UWB systems for communications and sensing applications in recent years. A variety of approaches have been adopted in both high- and low data rate applications, such as ad-hoc networking between devices, radar-imaging technology in medical applications and sensor networks with tracking and positioning capabilities [2,3].



Figure 1.1: Applications on UWB Technology.

Although different modulation schemes are applicable for UWB systems, one of the most popular is impulse radio ultra-wideband (IR-UWB). The IEEE 802.15.4a standard, that targets the Low-Rate Wireless Personal Area Networks (WPAN) describes a physical layer using the IR-UWB modulation scheme.

The high carrier frequency (between 3 to 10 GHz) allows the use of small form factor antennas with sufficient gain, reducing the overall form factor of the devices. For a UWB receiver both coherent and non-coherent detection scheme can be used, whose implementation is a tradeoff between performance and power consumption [4]. The nature of IR systems (i.e. the absent of a constant carrier tone) allows duty-cycling of transceiver, resulting in reduction of the power consumption and therefore increasing the battery lifetime of battery powered devices. These features give high potential to a world of new applications in wireless sensor networks (WSN) and body area networks (BAN). A comparison with two other communication protocols is made in Figure 1.2.

	2.4GHz ZigBee	2.4 GHz WiFi	UWB
Data rate	Low, 250kbps	High, 11 Mbps for 802.11b and 100+ Mbps for 802.11n	Medium, 1 Mbit/s mandatory, and up to 27Mbps for 802.15.4a
Transmission distance	Short, < 30 meters	Long, up to 100 meters	Short, < 30 meters
Location accuracy	Low, several meters	Low, several meters	High, < 50cm
Power consumption	Low, 20mW – 40mW	High, 500mW- 1W	Low, 30mW
Multipath performance	Poor	Poor	Good
Interference resilience	Low	Medium	High with high complexity receivers , low with simplest receivers
Interference to other systems	High	High	Low
Complexity and cost	Low	High	Low – medium – high are possible

Figure 1.2: Comparison between typical wireless standards [5]

1.1 Motivation

An ultra low power IR-UWB (impulse-radio ultra-wideband) solution for wireless streaming of high-quality music is developed by the Holst Centre. The system enables a music transmission over a distance of up to 20m and has the potential to be much more power efficient compared to traditional Bluetooth connection [6].

Since the targeted applications are battery-powered, a long battery life is essential. In addition to the concerning on the average power consumption, the peak power of the UWB radio solution should be limited since the high discharging current raises both the internal cell pressure and temperature in the battery. These aspects can quickly degrade the battery (reducing its lifetime) or even destroy it. For achieving a low power solution, the digital baseband architecture needs to be carefully optimized to reduce the total power consumption. In particular the compute-intensive synchronization phase, used to detect the start of an UWB data packet in noisy environment, needs to be approached with great care.

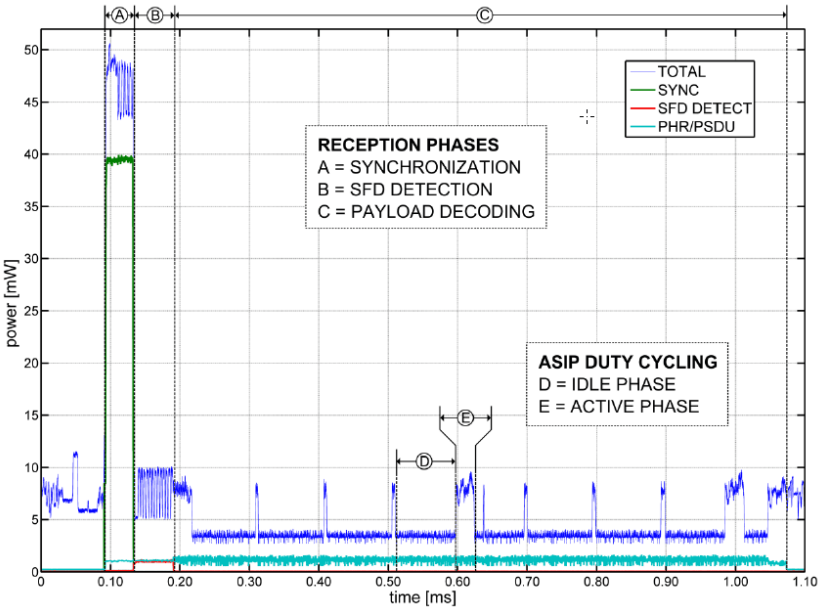


Figure 1.3: Power profiling of an UWB frame reception(source: Holst Centre/imec)

Figure 1.3 illustrates the power profiling result during an UWB frame reception. A high peak power during the timing synchronization phase can be observed. During this phase a matched filter is used to detect the start of the UWB frame and obtain the relative timing offset between transmitter and receiver.

Based on the above mentioned issues, a low power correlation operation in IR-UWB system is crucial. The fundamental motivation of this thesis is, to optimize the block that fulfills the task of synchronization, mainly focus on reducing the power consumption of the synchronization module, however, the hardware complexity and design area are also important.

1.2 Contribution

This thesis presents a systematic exploration of power reduction is performed in different level abstractions. More precisely:

- Different architectures are implemented and simulated for power concerning, the optimum one is further optimized to accommodate the UWB detection modes;
- Register and SRAM based circle buffer are compared as in the role of delay line, followed by register partitioning that is progressively explored to support several modes of operation.
- Reducing the switching activity in addition-rich module by changing the arithmetic number representation.

- Exploration of the achievable lowest supply voltage for shows that the voltage can be scaled from 0.84V to 0.7V, power simulation and analysis is iterated after cell re-characterized;
- The optimized design is benchmarked against a "standard" implementation, achieving up 2x power reduction with maintained throughput.

1.3 Outline of the work

In Chapter 2, an overview of ultra-wideband (UWB) background and synchronization for signal detection are given, moreover, the ASIC design flow together with low power methodologies are briefly discussed, a summary on the related work is included at the end of the chapter. Chapter 3 gives a overview of the achiteture and describes the procedure that the power optimization is explored in different abstractions. Chapter 4 presents the final evaluations in terms of power and area that were obtained from the various optimizations. Chapter 5 contains conclusions as well as the direction for future work.

2 Background

Ultra wideband radio (UWB) technology has attracted great interest for the use in wireless community in last decades. Typically, UWB systems can be categorized into high data-rate transmission (above 100 Mbps) and low data-rate transmission (around 250 Kbit/s) [7]. IEEE 802.15.4a is one of the standards that introduces new options for the physical layer in order to support low speed, lower power, better robustness for communication between nearby devices. These characteristics enable a variety of wireless applications like, wireless communication for medical devices, real time localization- and surveillance systems [5]. In the following section the background of the UWB will be briefly introduced. The synchronization methods used in UWB are described in the section 2.2. We present the digital design flow and low power methodologies in section 2.3 and 2.4 respectively. In the end of the chapter, a summary is made on the related works.

2.1 UWB principle

UWB is a communication method that originally designed for commercial radar systems. In 2002, The Federal Communications Commission (FCC) made the frequency bandwidth of 7.5 GHz available for UWB wireless communications. The interest was greatly magnified by the revision that has potential to meet the demand on convenient connection between devices in the short range wireless networks.

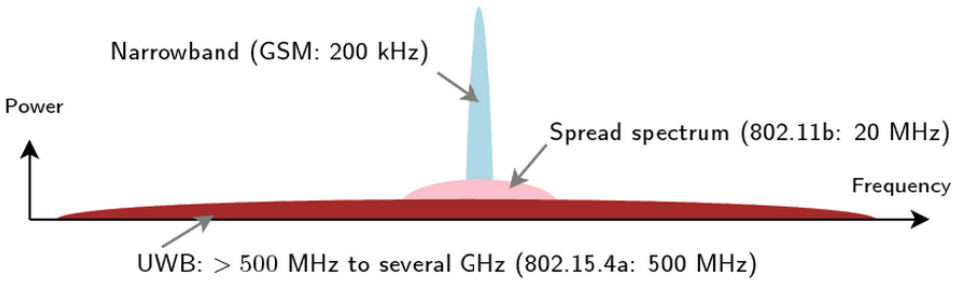


Figure 2.1: Ultra-wideband frequency principle

2.1.1 UWB definition

By definition, radio systems with a bandwidth larger than 500 MHz or the bandwidth larger than 20% of the carrier frequency are called UWB. The fractional bandwidth BW_f is defined as:

$$BW_{frac} = \frac{2(f_H - f_L)}{(f_H + f_L)} \tag{2.1}$$

where the upper F_H and the lower F_L refers the frequency measured at the -10 dB points to its center frequency.

Various frequency spectrum rulings exist over the world to minimizing the interference between different wireless communication systems. Currently, UWB systems are allowed to operate in the frequency band between 3.1 and 10.6 GHz, with a maximum allows power spectral density of -41.3 dBm/MHz.

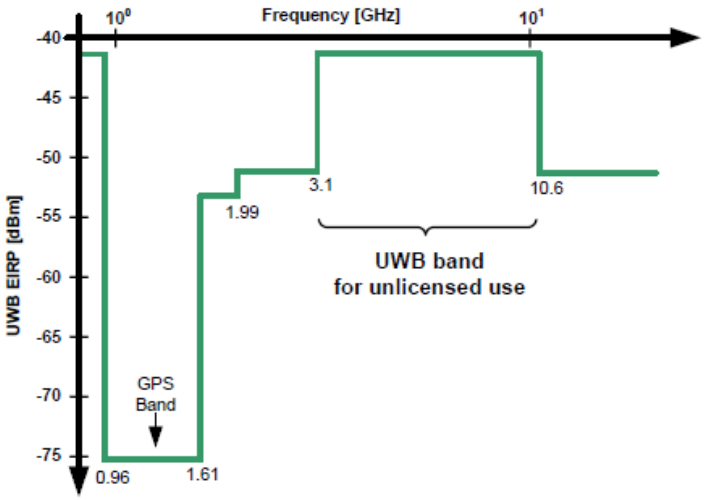


Figure 2.2: Ultra-wideband FCC spectral emission limits [8]

2.1.2 Impulse Radio

Impulse radio is a form of UWB communication using short pulses (around 2 ns duration). Multiple pulses are combined into a symbol by using a ternary sequence. Within this symbol the pulses of the sequence are separated by gaps where the transmitter isn't transmitting anything. This coding style is similar to other traditional direct-sequence spread spectrum (DSSS) since they all use short pulses. But there are a few fundamental differences between the two methods [9]. A comparison between a traditional narrowband wireless system and an UWB impulse radio system is shown in Figure 2.3. As narrowband signals are always present, so their duty cycle is 100 percent, while the transmitted pulses in UWB system have a low duty cycle. The property allows that the transceiver can be switched off when nothing is transmitted. This is a key feature to make the system power efficient.

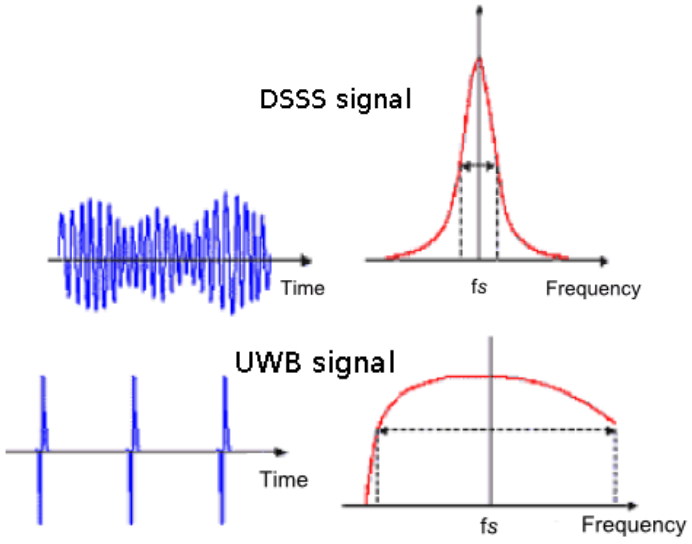


Figure 2.3: Duty cycle difference

2.1.3 Frame Structure

In IEEE 802.15.4a networks, the data is transmitted using the packet format which has three sections as illustrated in Figure 2.4.

- Synchronization Header (SHR)
  - Synchronization (SYNC) portion;
  - Start of frame delimiter (SFD) portion.
- Physical Header (PHR)
- PHY service data unit (PSDU)

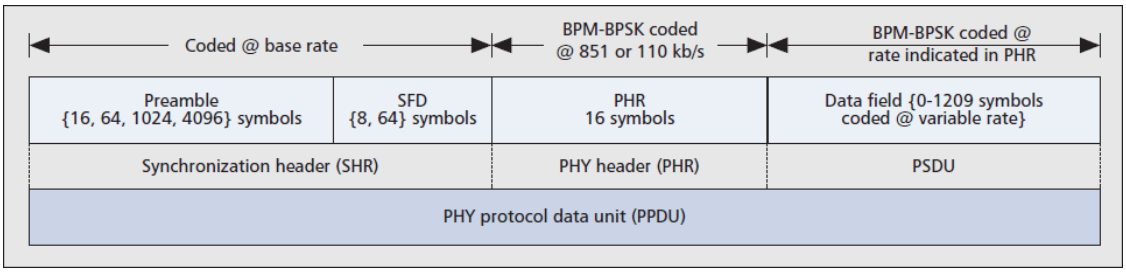


Figure 2.4: UWB PHY frame structure.

The synchronization header preamble is used to help the receiver to detect the packet, inquire the information on timing and channel estimation. The first part of the frame (i.e. sync portion), contains a sequence of concatenated preamble symbols with repetitions of a length from 16 to 4096 symbols. The second part of the synchronization header includes the SFD sequence to indicate the start of a payload.

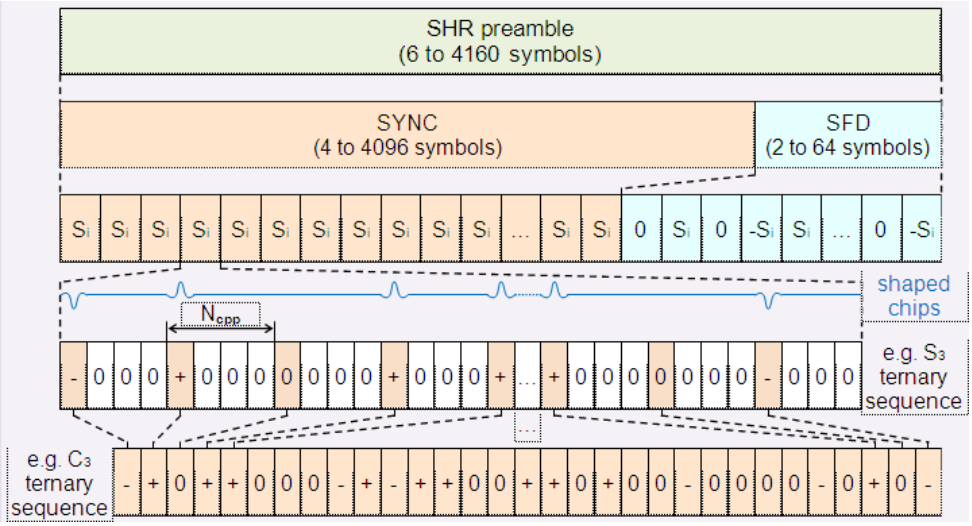


Figure 2.5: SHR symbols.

The Figure2.5 shows the detailed information contained in the SHR preamble. The preamble symbol consists out of successive ternary elements $\{-1, 0, 1\}$ which are separated by N_{cpp} empty chips. Depending on the operation mode the variable N_{cpp} can be 4, 16 or 64 chips.

The UWB PHY layer supports two types of length for the preamble symbols, i.e. 31 or 127 which can be used for a relatively high SNR and low SNR environment respectively. The preamble codes with a length of 31 are showed in Figure 2.6, the code index is used to identify the users when they use the multiple channels simultaneously. It can be observed that each code sequence contains a sequence of four consecutive zeros, thus shorter sequences (of 27 pulses) can be realized by realigning the code sequence [10]. This property can be used during the synchronization which will be explained in section 3.3.2.

Code index	Ternary sequence	Four zero aligned sequence
1	-0000+0-0+++0+-000+-+++00-+0-00	0000+0-0+++0+-000+-+++00-+0-00-
2	0+0+-0+0+000-++0-+---00+00++000	0000+0+-0+0+000-++0-+---00+00++
3	-+0++000-+-++00++0+00-0000-0+0-	0000-0+0--+0++000-+-++00++0+00-
4	0000+-00-00-++++0+-+000+0-0++0-	0000+-00-00-++++0+-+000+0-0++0-
5	-0+-00+++-+000-+0+++0-0+0000-00	0000-00-0+-00+++-+000-+0+++0-0+
6	++00+00---+-0++-000+0+0-+0+0000	0000++00+00---+-0++-000+0+0-+0+
7	+0000+-0+0+00+000+0++---0-+00-+	0000+-0+0+00+000+0++---0-+00-++
8	0+00-0-0++0000--+00-+0++-++0+00	0000--+00-+0++-++0+000+00-0-0++

Figure 2.6: Short preamble code sequences (source: Holst Centre/imec).

The following part after synchronization header is the PHY header. The PHY header contains information about the length and data rate of the packet. The PHY header data is protected by six parity check bits. In the PHY service data (PSDU), data is

coded using burst position modulation (BPM) and/or binary burst phase shift keying (BPSK).

2.2 UWB Synchronization

The receiver performs the channel estimation and demodulation of the data packets, where the signal synchronization is preceded by properly aligning the received signal with the template signal to provide accurate information on the reference phase and amplitude. An overview of a receiver is illustrated in Figure 2.7. The output of LNA are fed to a mixer to generate in phase and quadrature phase signals, followed by a variable gain amplifier(VGA) and Analog-to-Digital Converters(ADC), a digital baseband is used at the end for further data processing.

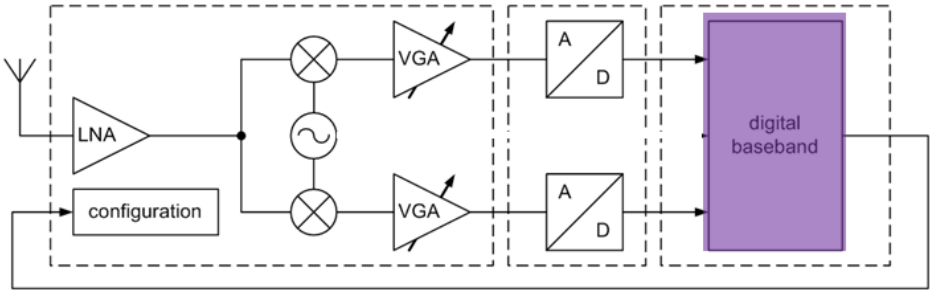


Figure 2.7: Overview of receiver in UWB system.

Unlike other communication protocols, the distinguishing features such as the nature of carrier-less, very low duty cycle operation and limited transmission power make the timing synchronization a difficult task when the receiver performs demodulation in UWB systems. By virtue of the energy of original symbols are spread over a wide bandwidth, the received energy level is likely below the noise level. Besides, such a wide bandwidth causes a fine resolution of the timing uncertainty region, thereby resulting in a large searching space size [11].

2.2.1 Synchronization procedure

Generally speaking, synchronization can be split into two parts, symbol code acquisition and fine tracking [11]. The purpose of coarse estimation is to determine the time when a symbol is detected in a short time, identifying the time instant as a reference boundary for the subsequent estimation. In the second step, the code tracking is a refine process of achieving a better alignment between the incoming and referenced pseudo-random noise (PN) sequences.

2.2.2 Types of preamble detection

The synchronization module can be operated in three different approaches: coherent, non-coherent and partially coherent. Coherent systems need to process both the in-

phase and quadrature signal, thereby requiring expensive and complex signal recovery circuit for a better bit error rate of detection. In the contrast, non-coherent systems discard the phase information and use methods like energy detection to recover the data. A hybrid scheme called partially coherent is proposed combing the two former mentioned methods [12]. In this scheme, the amount of coherent sets is chosen according to the degree of frequency offsets, thereby reducing the probability synchronization error in successive symbols.

2.2.3 Principle of correlation

A correlation process by estimating the degree of similarity between two streams is usually required in the area of communication systems. The phase of template is determined to best match with the input symbol. Therefore, the correlation based receivers are frequently applied due to is relatively better noise performance and robust for interference. Two kinds of correlation are classified: autocorrelation and cross correlation. Even though two functions of them look similar, there is a significant difference between them.

When two values of the same variable at times X_i and X_{i+k} are compared, the procedure refers to autocorrelation, as equation:

$$y[n] = \sum_{i=0}^{N-1}(x[i]x[i-k]) \tag{2.2}$$

The correlated signal can be seen as segments of the same received samples with different delays, the distance between correlation peaks suggests the fundamental frequency. The auto correlation based synchronization is very attractive because of this low complexity and power efficiency. However, it only performs well in the scenario where the signal-to-noise ratio (SNR) is high, otherwise the receiver may suffer performance degradation because the peak calculated by autocorrelation function of the current symbol might not indicate the correct timing instant.

In the environment where suffers from significant amounts of noise, cross correlation is preferred because it is an effective method of estimating the quality of the match between two independent signals to perform robust spectrum analyzing. The drawback of the method is that it requires complex hardware implementation compared with autocorrelation [14]. The corresponding formula can be expressed as,

$$y[n] = \sum_{i=0}^{N-1}(PN(i)x[i-k]) \tag{2.3}$$

where PN are coefficients represented by a pseudo-random pattern of chips. X are the input signal samples, N denotes the length of observation window and i is the discrete-time index. The Figure 2.8 demonstrates the procedure of cross correlation between the input signal and the template code.

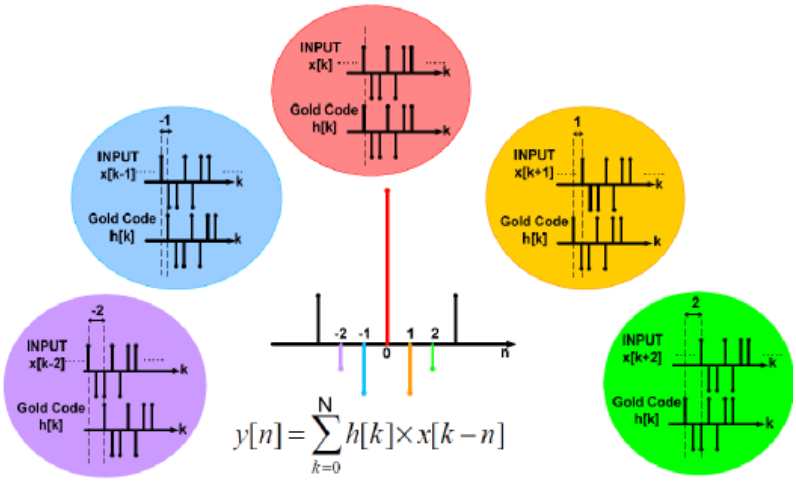


Figure 2.8: Spreading code phase detection.

The functionality is performed by a digital matched filter, as Figure 2.9 shows. The incoming signal is stored in a tapped-delay line and then multiplied with coefficients in each tap, which are a stored copy of the PN sequence. The results are finally summarized to the output. The magnitude of the value is computed and accumulated for a fixed numbers of times using an accumulator, then the position of max and min values will be used for detection.

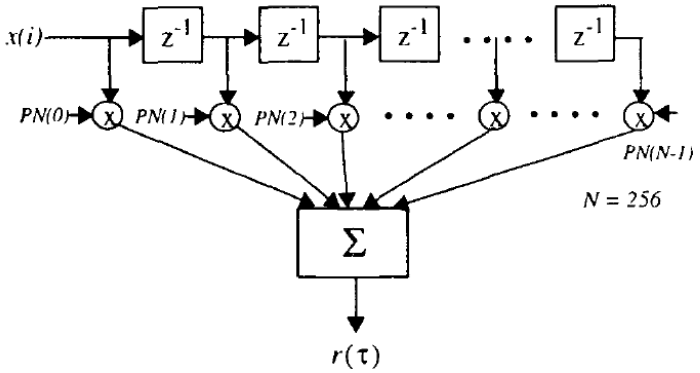


Figure 2.9: Structure of digital matched filter.

2.3 ASIC design flow

Figure 2.10 provides an overview of the application specic integrated circuit (ASIC) design flow, the procedure will be explained briefly in the following section.

2.3.1 RTL behavioral description and simulation

Hardware Description Language is firstly used to specify the behavioral description on Register Transistor Level (RTL), using the elementary circuit component to realize the

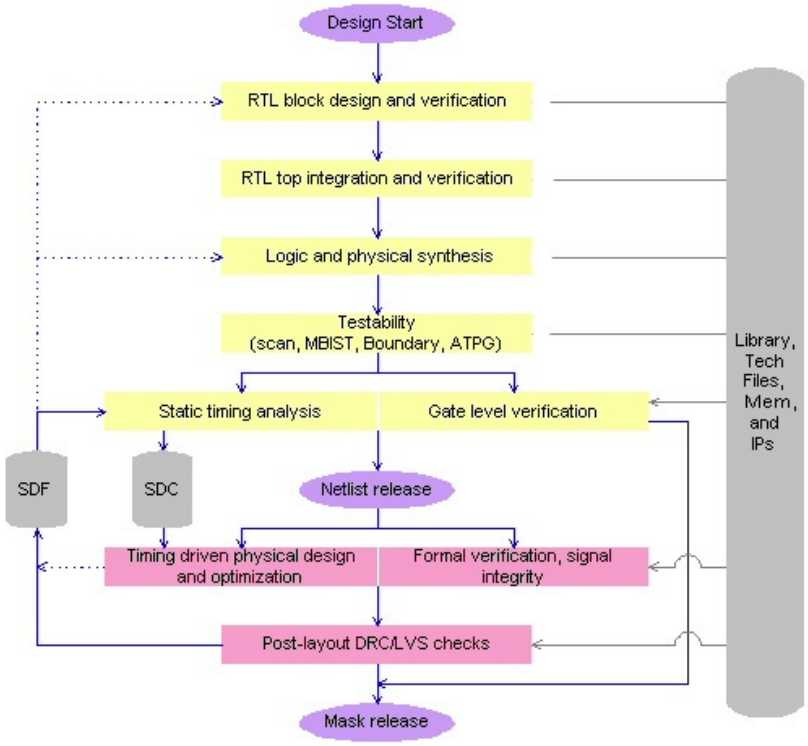


Figure 2.10: ASIC design flow.

specific functionalities. Together with a corresponding testbench, the RTL design is simulation to check whether the functionality is correctly defined.

2.3.2 Logic synthesis and simulation

Synthesis is the process of mapping RTL description onto standard cells provided by the technology provider (e.g. TSMC). In this process two stages can be distinguished:

- RTL Synthesis and Library Mapping: Given the specific circuit elements (i.e. standard cells), a logic synthesis tool transform the RTL description into a standard cell;
- Static Timing Analysis: The circuit is checked if all possible timing paths meet the timing constraints. Multiple iterations are executed to optimize the netlist until the timing constraints are satisdfied.

After the synthesis is done, a gate level simulation should be executed to check if the top level netlist with the technology libraries and SDF can achieve the same timing and functionality as RTL level.

2.3.3 Physical design.

During physical design, all the circuit components are placed and connected through routing in metal layers. The specific steps are described briefly as follows,

- Floorplanning: The dimensions of the ASIC core, the distance between core and in- and outputs (IOs) and location of the hard macros are defined. The modules are closely arranged to minimize the wire length;
- Power Routing: The core rows and the blocks are connected to the IO supply pads by introducing core and block rings;
- Placement: The gate-level netlist generated by logical synthesis is used to place the standard cells on the floorplan in this step. Various factors should be considered during this procedure, such as the total wire length, routing congestion and timing issue. All of these aspects affect the chips performance;
- Clock tree synthesis: The buffers are inserted to ensure all the clocked components are clocked at the same time;
- Design Routing: The design routing is performed to realize the metal connections between the pins of standard cells and macros specified in the imported gate-level netlist;
- Verification and generation: After the routing is completed, several verifications should be performed to check if everything is properly connected, including connectivity verification, geometry verification and timing verification. In the last step, we should export the related reports, and the final GDSII file.

2.4 Overview of low power design

In recent years, design closure in the ever complex devices requires a delicate balance of many issues, low power methodology attracts more and more attention because of the massive advent of portable devices. The increasing power densities on chips requires power efficient operation to ensure these portable devices can operate under severe power constraints. The total power dissipation for a CMOS design can be expressed in two types: dynamic power consumption and static power consumption [13].

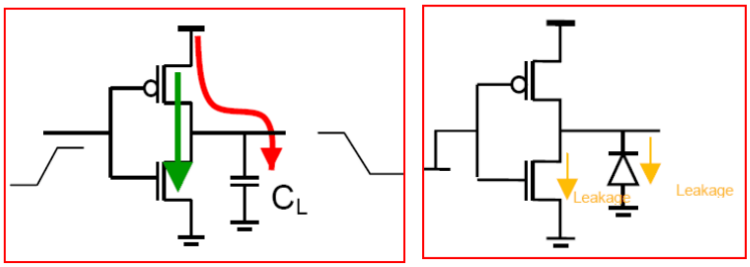


Figure 2.11: Dynamic power & Static power

Dynamic power consumption can be further grouped as:

- Switching power dissipation, indicated by the red line in Figure 2.11. As the name indicates, the power is dissipated during the CMOS circuit charges or discharges

of the output node capacitance.

$$P_{swith} = \alpha_T * C_{load} * V_{dd}^2 * f \tag{2.4}$$

where the total load capacitance C_{load} is the sum of the net and gate capacitances on the driving output, α_T represents activity factor, and the frequency (f) is the rate of state transitions.

- Internal power, indicated by the green line in Figure 2.11. The internal power is contributed by two factors. The first one is due to the energy consumed by the short-circuit current between the N and P transistors of a gate when both are on. One the other side, the transition of parasitic node capacitance associated with each internal node also dissipates the power.

$$P_{internal} = \alpha_T * C_{inter} * V_{DD}^2 * f + (V_{DD} * I_{sc}) \tag{2.5}$$

Static Power:
Static power has three basic sources, the dominant one comes from source-to-drain subthreshold leakage current when both NMOS and PMOS transistors are off. In addition to this, the reverse-biased of source and drain junction is another source of leakage, the percent of which is small. The last form is due to dc current in pseudo-NMOS when the output is low [13].

2.4.1 Low power design methodology

The ever-increasing demand for more functionality in smaller and more energy efficient portable devices motivates research on low power design solutions for battery-powered devices. With the revolution of semiconductor technology, low power in digital design depends heavily on the chosen process node. Furthermore, there are some other effective techniques for low power implementation: clock gating, power gating, dynamic voltage/frequency scaling, multiple threshold cells, etc during the design flow. Several optimization methods that achieve power goals are described in the following section.

- Clock gating: Clock gating is a common technique to reduce the power in clock distribution buffers and the associated downstream logic by reducing the activity. An enable signal is used to gate the clock of unused portions of the circuitry, hence reducing dynamic power dissipation;

- Lower the voltage: As the equation 2.4 indicates, it is easy to see that lower the power supply voltage is an effective technique to reduce both dynamic power and static power. However, it comes with the penalty of propagation delay in the transistors. The drawback can be compensated through decreasing the threshold voltage of the cells, enabling the circuit to achieve the same performance at a lower supply voltage. This however comes with the cost of increased leakage current;

- Power gating: The leakage in transistor is poised to contribute significant portion of total power for smaller technology nodes. Power gating is an effective method for alleviating this problem by temporarily shut down unused blocks;

- Multi-threshold devices: The technique refers cells with multiple threshold voltages Vth are used in circuit at the same time to make a tradeoff between delay and power. The high threshold cells are used concerning low leakage while the low threshold cells are placed on the critical path for their faster speed. The multi-threshold cells are made by different doping of the active area so that different gate characteristics are provided to meet the system requirement. However, the drawback of this technique is that it will increase the fabrication complexity;
- Dynamic frequency/voltage scaling: Dynamic frequency/voltage scaling is a technique to provide variable amount of energy for a task by automatically adjusting the frequency or operating voltage of a design on the fly. Even though the frequency reduction slows down the operation, just-enough computation is executed to meet the deadline of the task so that the power is consumed efficiently for lightly loaded machines.

A comparison is made between various power reduction methods, the effectiveness and impacts are summarized in Figure 2.12 [16].

Power Reduction Technique	Power Benefit	Timing Penalty	Area Penalty	Methodology Impact			
				Architecture	Design	Verification	Implementation
Multi-Vt Optimization	Medium	Little	Little	Low	Low	None	Low
Clock Gating	Medium	Little	Little	Low	Low	None	Low
Multi-Supply Voltage	Large	Some	Little	High	Medium	Low	Medium
Power Shutoff	Huge	Some	Some	High	High	High	High
Dynamic and Adaptive Voltage Frequency Scaling	Large	Some	Some	High	High	High	High
Substrate Biasing	Large	Some	Some	Medium	None	None	High

Figure 2.12: Power tradeoff between different low power techniques

2.4.2 Power analysis procedure

As power has become a critical part toward achieving design success, accurate power analysis is required for most of the VLSI designs. The power calculation methods can be generally categorized as static or simulation based [17]. The later one is preferred, because of its higher accuracy. For these simulations, a Value Change Dump (VCD) file is generated from a simulation recording the activity of all net in the design. The VCD file together with resistance and capacitance information (SPEF file) extracted from the design gives a more reliable estimate of power consumption. The VCD based power analysis flow is illustrated in Figure 2.13.

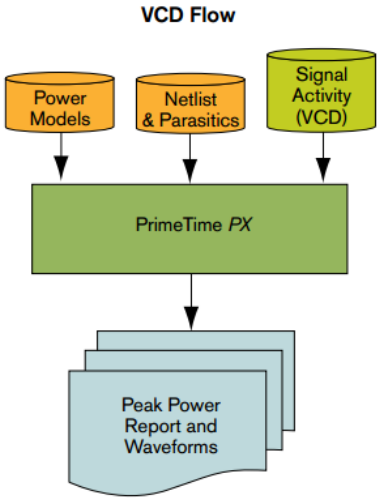


Figure 2.13: Power Analysis Flows [17]

2.5 Literature research

In the beginning of the thesis, a literature study on the state-of the art is performed to find the concepts for potential improvements. Most of presented correlator design for UWB systems are energy detection (ED) based [18]. In these systems, non-coherent receivers are used and the phase information contained in the samples is discarded to reduce the complexity and thereby reducing the power consumption. However, the current project requires full-, non-, and partially-coherent detection to adapt to different different transmission rate and SNR environment, hence a flexible scheme should be considered.

A sub-threshold baseband processor is designed in [19]. Massive parallelism is explored to achieve voltage scaling and less acquisition time without sacrificing the throughput. 620 correlators are implemented with the entire system operating under 0.4V, however, the leakage and interconnect capacitance are increased a lot due to the highly parallel architecture. Various methods, such as clock gating, custom layout and power gating are used to mitigate the influence of these problems.

On the other side, since UWB can be generally treated as an extension of traditional spread-spectrum systems, matched filters applied in these systems are also investigated.

A digital matched filter (DMF) is proposed in [20] for the application on Wideband-Code Division Multiple Access (W-CDMA). Multiphase clock generation is distributed in the reception registers, besides, correlation calculation operations are performed concurrently and a halting scheme is implemented in the threshold judgment.

An optimized design based on ripple-counter based correlator combining with biased representation is presented in [21] to reduce the power and the hardware complexity. The idea was further improved in [22], the authors further replaced the ripple-counter with custom-designed True Single Phase Clock (TSPC) registers and multiplexers, achieving 22.5% power saving on the incrementer.

In [23], the authors compared three methods on the realization of the correlator block. The result shows that the register file based correlator together with bus invert data encoding is the most power efficient architecture, data samples are stored statically while the single bit coefficients are shifted instead. The bypass adder tree can reduce a few computational expense and exhibits power advantage over the regular adder tree, but not fully compatible with the register file based architecture.

3 Low power Exploration

The power consumption is becoming a crucial issue for a portable system design in recent years and drawing more research on the methodologies on reducing the power dissipation of integrated circuits. Generally speaking, power reduction can be realized at different levels of design abstraction (i.e. system-, architecture-, RTL/logic- and the circuit level [24]). An optimized algorithm can effectively reduce the complexity to solve a specific problem and enable the low power design. At the system level, reconfigurable computing can be further developed by partitioning the algorithm between software and hardware implementation. At the architectural level, parallelism and pipelining allow a reduction in supply voltage without degrading system throughput. Clock gating can be applied either at architecture- or gate level depending on the specific application. Various data representation and invert bus coding are commonly used at the RTL level. A variety of design techniques, such as advanced technology process and Multi Vt can be used at the circuit level to achieve power minimization. The Figure 3.1 illustrates the opportunities provided for power reduction at different levels of design abstraction.

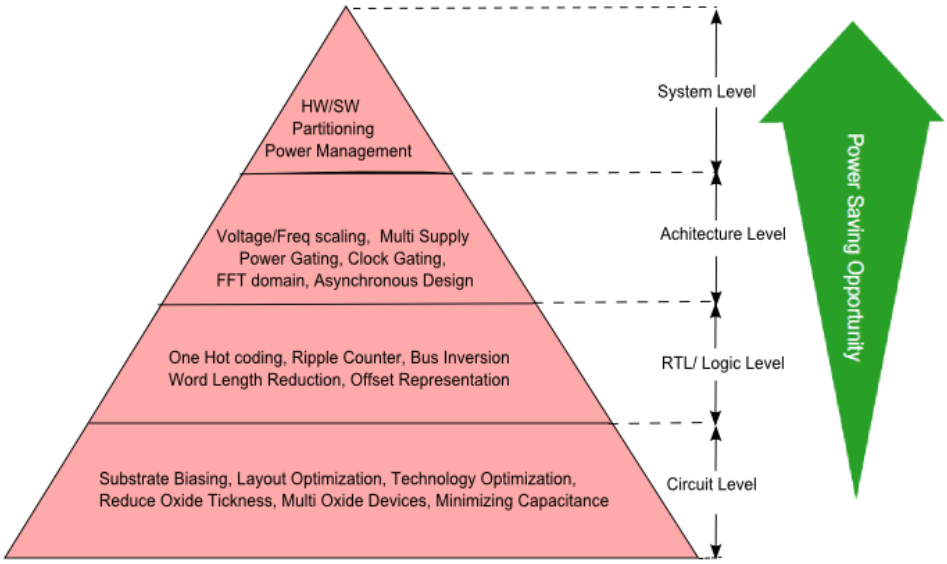


Figure 3.1: Levels of design abstraction for power saving [25]

In the next subsections the the method to perform preamble detection will be described. Afterwards, the evaluated and applied low power techniques will be explored to reduce the power consumption of the synchronization module.

3.1 Preamble detection procedure

In the UWB-IR system, preamble detection is required to determine the presence and timing position of the preamble. A general symbol-to-symbol correlation process can be modeled as Figure 3.2. The result provides an improved output signal SNR for preamble detection.

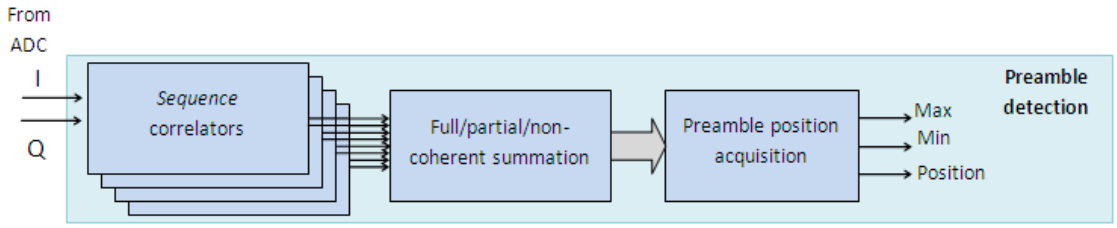


Figure 3.2: Preamble detection flow

A matched filter is used to measure similarity between a known code sequence and the received baseband samples. The inputs and output of the matched filter are complex signals, in the next step the two parts are combined by magnitude calculation and summation in different manners according to the chosen detection mode. The accumulator module inside the timing synchronization accumulates the correlation results over multiple preamble symbols and extracts information about the maximum value, minimum value and maximum position. From the accumulation result the maximum and minimum value are extract together with the position of the maximum value [26]. This process is illustrated in the Figure 3.3 below.

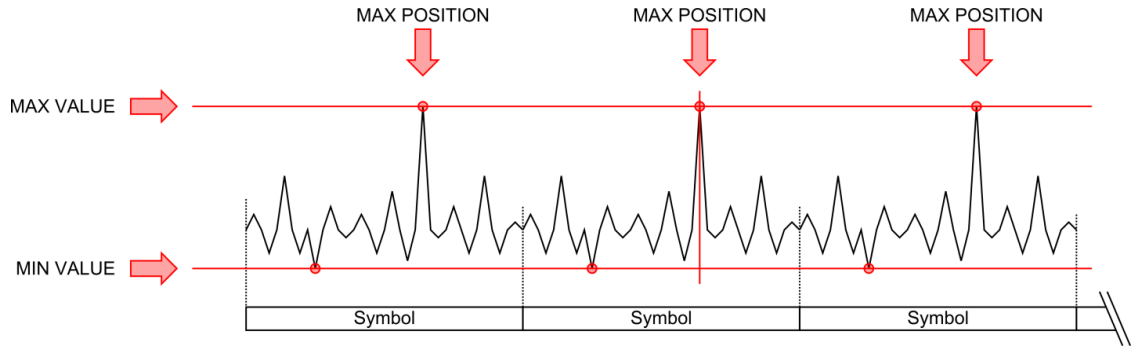


Figure 3.3: Position determination process of the accumulator(source: Holst Centre/imec)

These three parameters are updated after N symbols (where N equals the number of symbols accumulated) and passed to the ASIP core for further processing.

In order to adapt to different noise environments and support different modes from the IEEE802.15.4a standard, the preamble detection module contains several configuration signals:

- Mode: Select the coherency level of the synchronization module. This two bit value is used to select between the following three modes:

- full-coherent
- quarter-coherent
- non-coherent

- Offset: Since Ncpp-1 empty chips inserted between successive ternary preamble symbols in the IEEE802.15.4a standard, the parameter specifies the offset which is the isolated pulses for different spreading factor (i.e. 4, 16 and 64). The setting is used for compensating the delay between the partial correlation results of the sub-correlator.

3.2 Architecture optimization

One common way to implement a correlator is through a digital matched filter (DMF), either in the strcuture of direct form or transposed form. Assume the sample rate is equal to the chip rate, and the expression of an Order-N digital matched filter is given by:

$$y[x] = \sum_{i=0}^{N-1} (s[(M * i) * x] * C_i) \tag{3.1}$$

Where S[n] is input sample for the digital matched filter, M represents the spacing between the isolated pulses. C[i] is i-th coefficient of the spreading sequence, which is a ternary sequence for PN code. N represents number of taps of the DMF system. The Figure 3.4 shows the diagram of the direct form structure for the hardware implementation.

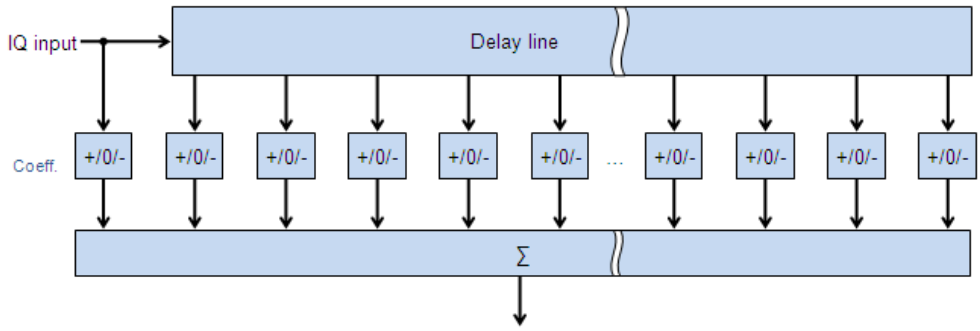


Figure 3.4: Direct form of preamble correlator

By assuming that the input variables s[n] are stored in memory, the equation 3.1 shows that the (read) access rate is directly proportional to the number of taps. Therefore a power efficient solution of storing the x[n] values becomes crucial for a low power implementation. The length of the PN sequence is a variable value in the UWB system. The issue will become more serious when the length increases, no matter the delay units are implemented in the form of memories or registers. Besides, massive hardware resources will be consumed on multipliers and additions in the straightforward structure. For example, the final summation requires significant numbers of adders so that the critical path is also directly proportional to the number of taps. Generally speaking,

the complexity of the correlator implementation depends on two main components: the length of delay units and the number of multipliers and adders. Therefore, direct form structure of the DMF should be optimized to mitigate the power dissipation caused by the above mentioned facts.

3.2.1 Architecture comparison

Since the expression of DMF looks similar to the common Finite-length Impulse Response (FIR) filters, a power comparison is made between the direct form and segmented form based FIR filters. A generic design is implemented in which the taps varies from 16 to 256. The input samples are 5 bits two's complement values, equivalent to the targeted application. The length of the coefficients equals to the numbers of taps, each of which is 7 bits. 16-tap direct and segmented realizations are shown in Figure 3.5 (a) and (b) respectively. The segmentation factor is determined to four in this experiment.

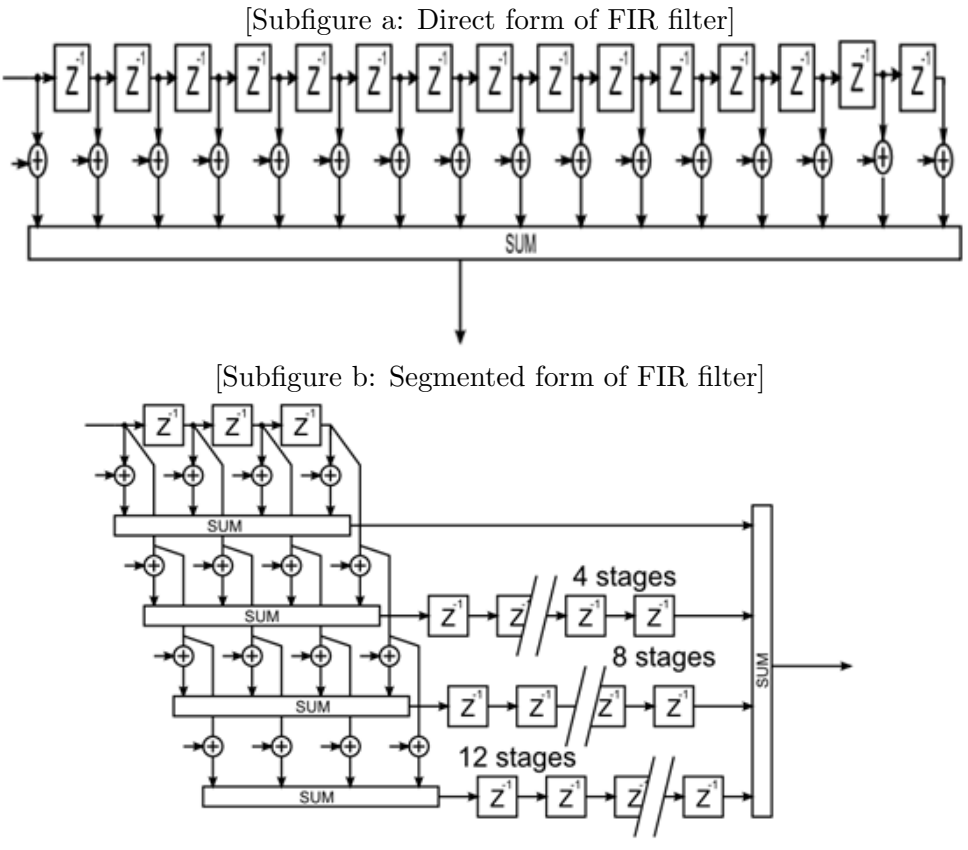


Figure 3.5: Two kinds of FIR achitecutre

In the segmented architecture, the common FIR design is changed from 1-dimension structure to a 2-dimension one. The tapped delay lines is shortened by N times and shared by each sub-filter. Likewise, the coefficients are also divided into several sub-

sections. The equation representing this implementation is:

$$y(x) = \sum_{i=0}^{W-1} s[i*x]*C_{seq0} + \sum_{i=W}^{2W-1} s[W+i*x]*C_{seq1} + \sum_{i=2W}^{3W-1} s[2W+i*x]*C_{seq2} + \sum_{i=3W}^{4W-1} s[3W+i*x]*C_{seq3} \tag{3.2}$$

Where W is the length of shared tapped delay line, the coefficients in each subset are partitioned as:
$C_{seq0} = C_i \quad fori = [0.....3]; C_{seq1} = C_i \quad fori = [4.....7];$
$C_{seq2} = C_i \quad fori = [8.....11]; C_{seq3} = C_i \quad fori = [12.....15];$
Equation 3.3 can be transferred to the z-domain:

$$y(z) = S(z) * C_{seq0} + z^W * S(z) * C_{seq1} + z^{2W} * S(z) * C_{seq2} + z^{3W} * S(z) * C_{seq3} \tag{3.3}$$

The equation shows that the separate outputs from the sub-filters have to be delayed in time. The length of these extra elements are multiple times as that in the shared tapped delay line, consequently resulting in more registers(memory) cells in the design after the architecture modification. In order to mitigate the overhead, the formula can be further mathematically transformed to Equation 3.4, which means the output after each sub-filter can share a single serial delay line for the power and area concerning.

$$y[z] = S(z) * C_{seq0} + z^W \Big\{ S(z) * C_{seq1} + z^W * [S(z) * C_{seq2} + z^W * S(z) * C_{seq3}] \Big\} \tag{3.4}$$

In contrast to the parallel one, the delay elements used after the sub-filters have an equal length and shared by the each sub-correlator in cascade. The output of the i th delay line is added to the output of the (i-1)th sub-filter and fed to the following one. In the end, the final correlation values can be obtained after (i-1) delay elements and adders, as Figure 3.6 illustrates.

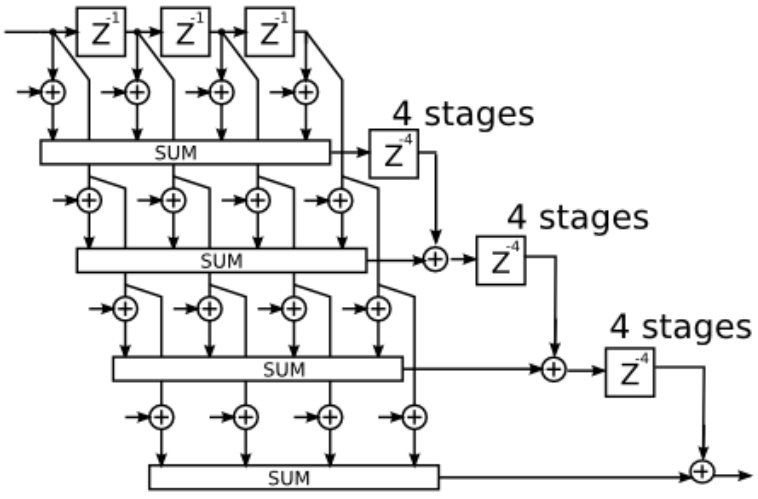


Figure 3.6: Segmented form of FIR filter with serial delay lines

Even though both the memory access rate and adder tree will be reduced as analyzed before, the extra delay lines are still a potential factor that might have negative influence

on power consumption. It is important to choose an optimum value for segmentation, since it is a critical factor that determines both the access time and numbers of delay line. Therefore, a comparison is performed based on the above mentioned FIR filter. The result, as illustrated in section 4.2.3, indicates that the segmented form with serial delay line becomes power efficient than the direct form when the tap increases to 128 and over. Based on this result, the segmented architecture is adopted in the correlator design. The reasons rely on two facts:

- Incoming samples from the ADC are 4-channel data in a complex quantity consisting of an in-phase part and a quadrature part, and the length of preamble codes is 31, which means 4*2*31=248 taps will be used in the direct form.

- Since the signal is sampled at about 500MHz, by optimizing to the four segmented architecture, the memory access rate is reduced from 500M*31= 15.5G access/sec to 500M*(6+3) = 4.5G access/sec when compared with the direct form, almost one third of the naive correlator rate.

3.2.2 Overview of the optimized architecture

Using the notation from above, the architecture used for the preamble detection is illustrated in Figure 3.7. The correlator design can be generally categorized to two sub-modules (i.e. *subcorr* and *delcorr*).

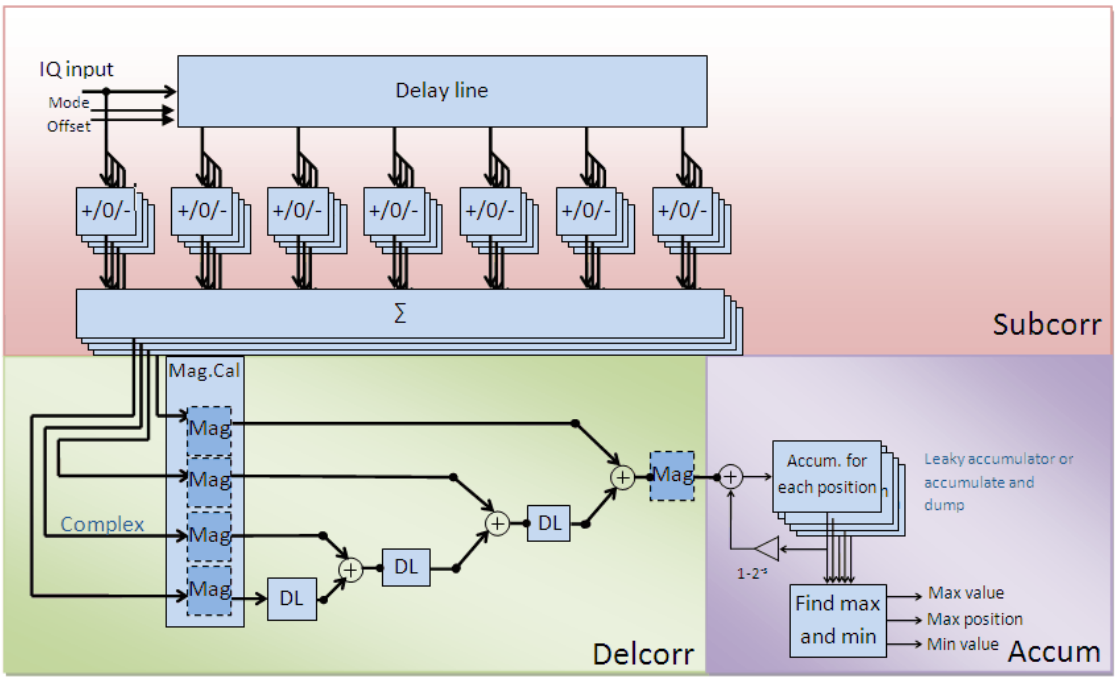


Figure 3.7: Overview of correlator architecture

These blocks are described in more detail in the following subsections.

- *Subcorr*: implemented as delay elements followed by multiply-accumulator components.

In chapter 2, it was shown that the maximum length of the PN code can be reduced to 27 non-zero elements after the realignment of the sequence. In order to be able to support different levels of coherency this PN code is split into four subsequences, allowing quarter coherent detection. As mentioned above, a segmented FIR structure is preferred over the direct form implementation, because of power efficiency. The *subcorr* module therefore contains a seven taps delay line used to calculate the 'sub-correlations' of quarter preamble symbols. The values stored in the seven tap delay line are therefore multiplied with four sets of ternary coefficients. It is important to mention that real multipliers aren't needed due to the ternary (+1, -1, 0) codes sequence. Instead, a three way multiplexer can be used to select the input signal, the negative version of the input signal or zero depending on the coefficient. In the design, the coefficient is represented by two signal 'enable' and 'polarity' as shown in table 3.1.

Coefficient	label Polarity	label Magnitude	Signal
0	0	0	0
+1	0	1	S[n]
0	1	0	0
-1	1	1	-S[n]

Table 3.1: Representation of ternary multiplication in *subcorr* module

- *Delcorr*: delay and accumulate the separate outputs from the sub-correlators, perform the magnitude calculation according to different modes.

  The module is designed to allow different levels of coherency for the synchronization. None-, quarter- or full coherent operation can be chosen flexibly by changing the order of addition and magnitude calculation. The separate outputs from the sub-correlators are delayed in time before the final summation. The detailed data path will be explained in the following subsection.

Finally, an accumulator is included to be able to accumulate over multiple preambles and determine the maximum value and position. For completeness this module is shown in the block diagram 3.7, however the optimization of this block was not part of scope of this thesis.

3.2.3 Data path optimization in *delcorr* module

The data path in *delcorr* module is optimized to perform sub-sequence delay and magnitude calculation depending on the operating mode. The signal data flow from the I- and Q-channel is depicted in Figure 3.8.

In case of non-coherent detection, the phase information is ignored and only one of the two channels (i.e. the I channel) is used for magnitude estimation. In this mode of operation the data path of Q signal can be isolated and clock gated. The clock gating technique thereby reduces the activity on the clock tree and its endpoints (i.e.

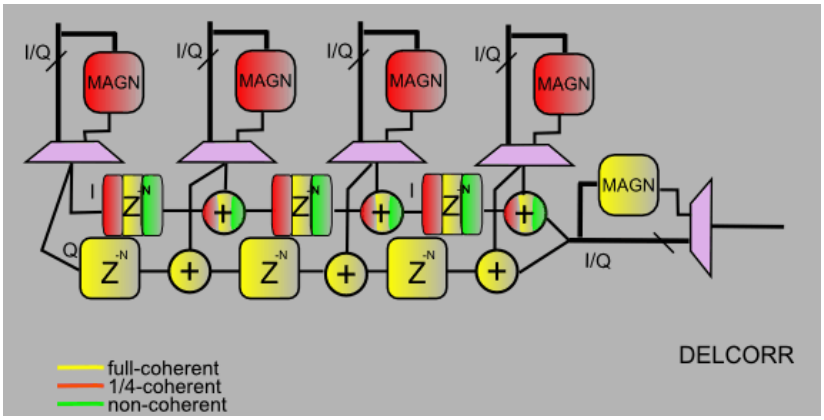


Figure 3.8: Detailed data path for summation and magnitude calculation

the flip-flops). Doing so, nearly half of the power consumption caused by delay units in *delcorr* module can be saved, during non-coherent mode. The active blocks used in the data path during non-coherent mode are indicated with the green color in Figure 3.8.

For quarter coherent detection both I- and Q channel are used in the sub-correlator. The results from the sub-correlations are passed through the magnitude calculation modules. After the magnitude calculation modules only one data path is needed, similar to the non-coherent mode. Therefore, also for this mode the Q-channel can be isolated a clock gated reducing the power consumption significantly. In Figure 3.8 the active blocks used for quarter coherent detection are highlighted in red.

For the full coherent detection case, the results of all sub-correlations are combined together while still using the complex values. The final result is passed through a magnitude module calculating the magnitude value. During this mode of operation the final magnitude block and all the delay line units from both the I- and Q-channel are used. In Figure 3.8 these blocks are highlighted in yellow.

3.3 Design of sub-sequence delay line

After the fundamental architecture has been determined, the next step is to explore the possibility of power optimization of the sub-modules. Power simulations identify the delay unit as the biggest power contributors in the proposed architecture. In the original design, these delay lines are implemented using static random access memories (SRAM). We firstly investigate the power variation by changing the access pattern of SRAM. After that, a comparison is made between SRAM based and register based circular buffer in this specific design. Lastly, we introduce the power reduction obtained by using register partitioning, which combines clock gating and operand isolation to minimize the dynamic power in the circular buffer.

3.3.1 Access pattern changing on SRAM

Delay lines are widely used in many applications to provide multiple clock cycles of delays. It is usually implemented as circular buffer for power consideration [27]. Instead of moving the data itself, pointers are moved through the address range of the circular buffer. Every clock cycle data is written into the location pointed by the address pointer. Typically, the write pointer is initialized at the beginning of the circular buffer and then moved to access the consecutive values. Meanwhile, the read index is following the write pointer according with a certain offsets to the write pointer. This access pattern is illustrated in Figure 3.9, the maximum size of delay line is determined to N. The write- and read index is initialized to 0 and M respectively. The distance between read index and maximum value, which equals to N-M in this case, is the delay time used in the circular buffer. In this manner, whenever the pointers reach the last location, they wrap back to the beginning of buffer and the newest data overwrites the previous data.

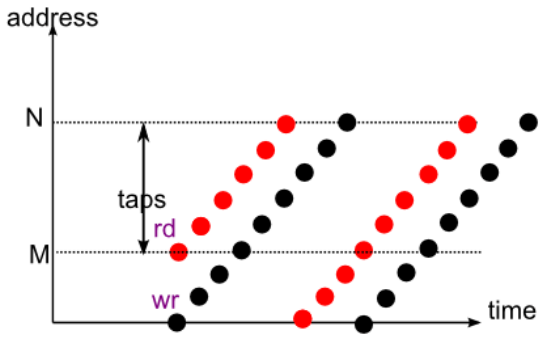


Figure 3.9: Memory access pattern based on two pointers

The access pattern for this circular buffer is very regular, so only one pointer is enough to read and write the elements in a sequential order. After n clock cycles, the data written into the memory is read out of the buffer while new data is written to the same location. Figure 3.10 illustrates the idea of using only one pointer as index. This modification reduces the activity of the address bus and therefore has the potential to minimize the power consumption of the memories. The results are shown in section 4.2.1.

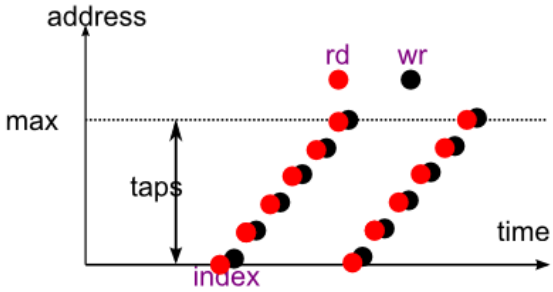


Figure 3.10: Memory access pattern based on one pointer

3.3.2 Comparison on Flip-flops and SRAM based circular buffer

The circular buffer can be implemented by either registers or static SRAM. In contrast to SRAMs are often used for big capacities, the flip-flops based implementation becomes more area and power efficient for small sized circular buffers. This is mainly due to the fact that the circuits overhead in memories such as sense amplifiers and address decoders. For this reason, a comparison based on power simulation is executed to determine the optimal choice for a given buffer size in this application.

3.3.3 Register partitioning

Generally speaking, the traditional flip-flops based buffers have three major components of power consumption:

- The data inputs of all registers are connected to the data bus;
- The logic which is used to control the data flow;
- The power due to the clock tree.

It is important to know, that the power caused by the transitions at data input to the flip-flops contributes the most significant portion. The switching activities caused by the pointer access consumes power dissipation along the data path of all flip-flops in each delay module, however, most of the flip-flops are kept in idle state. Hence, clock gating and operand isolation are used to decrease this portion.

The clock gating cells are used to save dynamic power by cutting off the clock distribution to a particular group of flip-flops in cases they don't need to be updated. In addition to the automatically inserted clock gates during the synthesis stage, module level clock gating, e.g. library specific Integrated Clock Gating (ICG) can be instantiated manually in the RTL code to gate the clocks of specific registers. A clock gating circuitry from TSMC vendor library is used in our design.

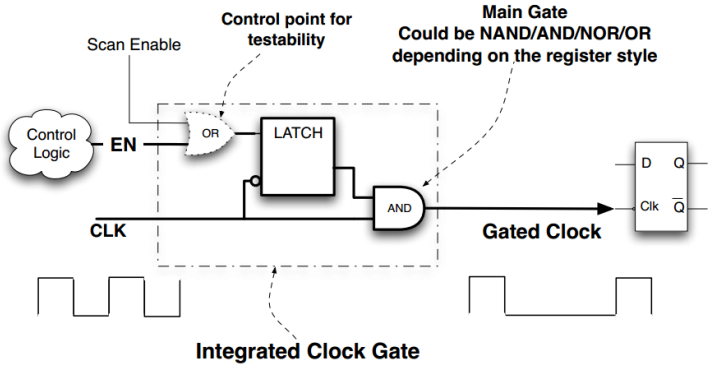


Figure 3.11: Integrated clock gating circuit [28]

In the current correlator design, the length of the delay line is determined according to spreading factor in the system, i.e. 4, 16 or 64, as defined in the IEEE802.15.4a standard. Therefore, the maximum length of the delay line has to be set to L=64/4*7=112.

However, in case of 4 or 16 chips/pulse only a fraction of the memory is used. Figure 3.12 shows that only a small portion (7 addresses) of the total memory size will be used in case of 4 chips/pulse. In the contrast, the total block (112 addresses) will be sequentially accessed when the spreading factor is set to 64. Consequently, the power is significantly consumed in this senerio.



Figure 3.12: Memory usage in sub-delay line

In order to keep a balance on the power dissipation in different cases, the memory space of the circular buffer is equally partitioned. From the address generation unit that control signals are derived to activate the partitions depending on the current address of the read/write pointer. Figure 3.13 shows a block diagram of the proposed partitioning scheme.

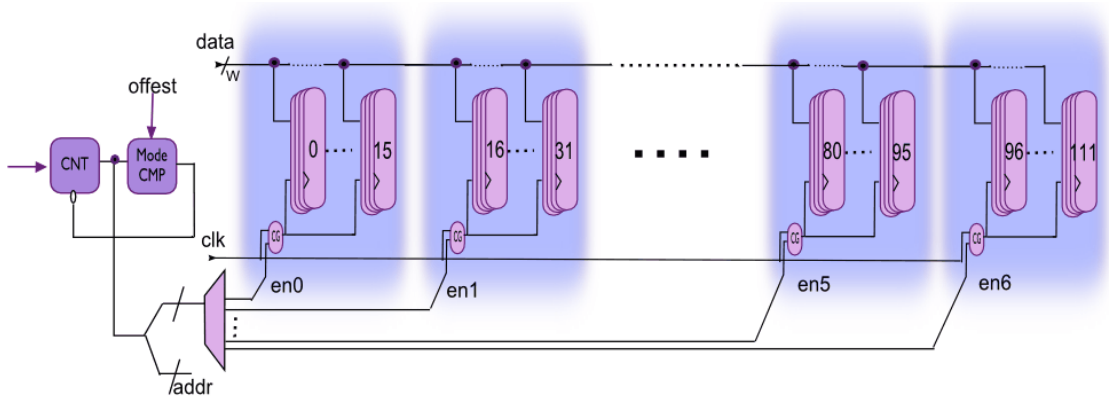


Figure 3.13: Partitioned registers with clock gating

The register based memory is split into 7 subsections of 16 bits address spaces. The three MSBs of the address pointer are decoded using a one-hot decoder, generating seven enable signals used to control the clock gates for the sub-registers.

Operand isolation is another widely used technique to reduce the activity in a digital design. The inputs signals of unused parts of the design are gated when idle. Data propagation into the module is thereby avoided by a control signal. Operand isolation is often used in combination with clock gating. Combining both techniques assures that there is no activity in the idle modules, thereby reducing the power consumption to the leakage power.

The concept of operand isolation is explored in the circular buffers use in the design. As shown in Figure 3.14, based on the existing partitioned circular buffer, AND gates are inserted as isolation elements. The enable signals used for the sub-modules are derived from the address pointer. Operant isolation ensures that the samples on the data bus are only passed to the enabled partition.

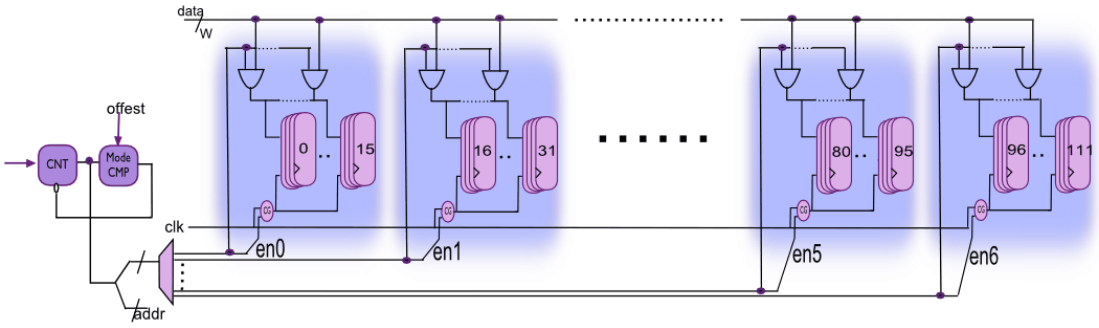


Figure 3.14: Partitioned registers with clock gating and operand isolation

Theoretically, the register partitioning can be progressively explored, which means we divide the delay line into more segments. With smaller size in the each partition, data transitions in sub-registers can be further reduced. However, although the use of register partitioning is an effective way to reduce the activity and thereby minimize the power of a design, its important to keep in mind that the extra logic needed for operand isolating consumes additional power and leads to area increase. In practice, it will be shown in section 4.2.4, the introduced overhead is limited compared to power savings from this technique. Hence, effective use of this technique can result in substantial dynamic power reduction. Another cautionary note in this design is that, it will make the multiplexer become more complex and is hard to meet the timing constraints when more partitions are inserted, especially under the low voltage circumstance. Based on the balance between timing issue and power consideration, the partition number is finally determined to 7.

3.4 Arithmetic Level exploration

Two's complement number representation is particularly suitable for the implementation of arithmetic operations, especially for the multiply and accumulate function. The reason replies on the fact that the addition and subtraction circuitry dont need to examine the signs of its operands.

Even though the computation efficiency brought by two's complement number, arithmetic operations such as addition, subtraction and multiplication requires a sign extension for every operand by appending digits to the MSB bit. The extension induces more capacitive load and undesired transition during the computation, more power will be consumed in these cases. To avoid sign extension, both sign magnitude and biased representation have been investigated in [29,30], comparing area and power consumption in theoretical and practical experiments. The results of the studies show that the biased number representation has power advantage over the conventional two's complement approach, and slightly better than the sign-magnitude approach. In term of benefit on area measured in transistors count, the architecture based biased number representation is a more desirable choice over the other ones.

The principle of biased representation is coding the signed numbers into unsigned numbers by adding a positive value to the input samples, as Equation 3.5 describes,

$$X = x_0 + \sum_{i=0}^{W-2} (x_i * 2^i) \tag{3.5}$$

By virtue of advantage on both area and power, biased representation is used in each sub-correlator module since there are massive accumulation operations. However, a straightforward two's complement representation was employed through the whole module and magnitude calculations are required in the *delcorr* block, two conventions are needed before and after the multiply-accumulate module. The original 5-bit signed data stored in each tapped delay line has a range from -16 to 15, and then 5-bit positive values are generated with a range from 0 to 31 by inverting the sign bit of original data. Due to the convenience in positive value addition, the sign extension is avoided in multi-operand summation to save the capacitance and the nodes activity, thereby reducing both and power dissipation significantly.

3.5 Circuit Level exploration

As equation 2.4 indicates, voltage scaling offers a very effective way to reduce the total power consumption, because of its quadratic relationship to power. However, such a voltage reduction results in a penalty that the cell delay increases significantly. This becomes a real problem when the voltage supply approaches the threshold voltage (V_t) of the devices.

Parallel and pipelined architectures are widely used to compensate for the increased cell delays at reduced supply voltages to meet throughput constraints. However, area boundary exists and limits the scalability. Alternatively, the increase cell delay at lower supply voltages can be compensated by changing the threshold voltage of the used technology. Most foundries provide standard cells with different threshold voltages. The 90 nm library from TSMC used for this design provides three threshold voltages (i.e. low, regular or high V_t). Although, decreasing the threshold voltage has a positive aspect on the cell delay, it has a negative impact on the leakage. However, the increased leakage voltage can be compensated by power gating techniques.

The initial synthesis is carried out using multi threshold cells. The CAD tools (i.e. Cadence RTL compiler) automatically chose between high V_t and normal V_t library depending on the timing requirements. The optimum V_t cell is determined based on a trade-off between the speed requirements and control of the leakage power. However, since the synchronization module of the receiver is only active during a faction of the frame length, the module is power gated during the off-periods reducing the influence of the leakage power.

Then the second iteration of synthesis is performed based on low V_t library for the timing consideration. To explore the achievable lowest supply voltage, the clock

frequency is iteratively increase to its maximum value of 200 MHz This increase in frequency can be translated into a decrease in voltage by using Equation 3.6 [31].

$$T_{delay} = K * \frac{V}{(V - V_{th})^2} \tag{3.6}$$

The TSMC datasheet shows, the threshold is 0.4V in this case. As described earlier, the delay is shortened from 8 ns to 5 ns due to low V_t cells, thus enabling the supply voltage can be reduced from 0.84V to 0.7V by theoretical calculation and results in approximately 30% power reduction.

In practice, we find the critical path between the circular buffer module, as highlighted in red line in Figure 3.15, that doesn't meet the timing constrains when we re-characterize the cell library to the 0.7 Volt versions. Hence, to compensate the loss of performance, pipeline registers are inserted after the second multiplexer to shorten the critical path.

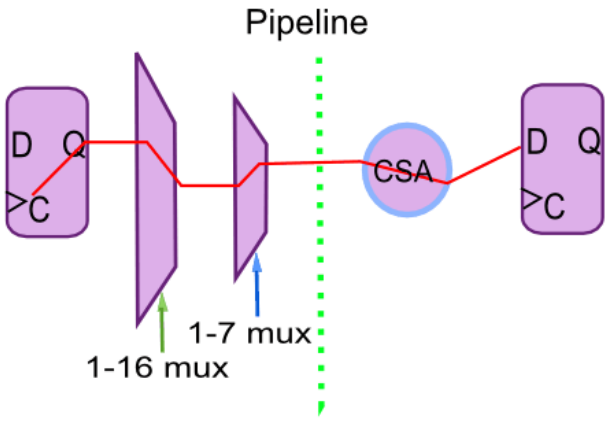


Figure 3.15: Critical path in the correlator under 0.7V cell library

Evaluation and Experimental Results

4

Based on the design considerations discussed in previous chapters, in this chapter the results of various evaluations on the proposed optimization methods are analyzed. These experiments are operated using full-coherent detection mode. Secondly, the power results of a back annotated netlist after synthesis and Place and Route are presented, using different detection modes. Finally, the overview of physical layout is shown.

4.1 Evaluation on power optimizations

In this subsection, the detailed comparison results and power improvements for the different optimization methods are provided. All the results are based on post-layout power simulation. The power simulations are preformed using Synopsys Primetime. The activity of the nets is derived from a VCD file which is extracted from a simulation using Cadence IUS. The used standard cells are provided by a 90 nm technology library of TSMC. If not explicitly mentioned, all results are assuming that the standard cells operate at 840mV while the macros cells (SRAM) are operating at 1.2V.

4.1.1 Access pattern change

Based on the original design, the experiment is done by changing the access pattern to reduce the activity on the address bus of the memories. The results in Table 4.1 show that the overall power reduction is less than 1%. After further analyzing the changes in memory sub-module, it can be found that the improvement by this changing is very limited and can be negligible. In fact, even though power consumed due to the toggle of all bits of address bus cause almost 20% of total memory power in typical case, the sequential access contribute little switching activities on the address bus.

Power Consumption				
Module name	modified		original	
	Dealy buffer(mW)	Memory(mW)	Delay buffer(mW)	Memory(mW)
mems2	4.65	3.83	4.68	3.81
mems3	4.24	3.72	4.28	3.73
mems4	4.62	3.97	4.63	3.99
total	13.51	11.552	13.59	11.53

Table 4.1: Power comparison between the two access patterns

4.1.2 SRAM and register based circular buffer

Initially, the sub-sequence delay lines are based on SRAM and parallel designed. Figure 4.1 illustrates the results on the process of replacing the three SRAM with registers step by step. For the smallest size memory (i.e. 128words x 128bits) the replacement with registers can get some power savings. However, there is no improvement for the two larger ones (i.e. 256x128 and 384x128 bits). From the area view, it is also not an advisable choice to do so, since the total area (after synthesis) will increase from 701524 um^2 to 1866383 um^2. This is an increase of more than 2x.

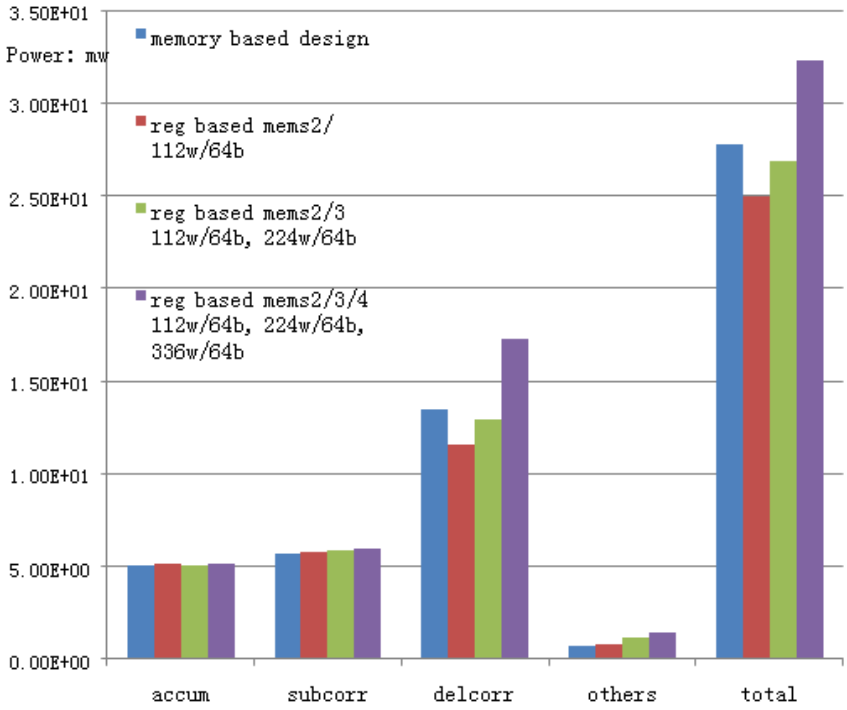


Figure 4.1: Power comparison between SRAM and DFFs

4.1.3 Architecture optimization

The three different architectures of 16 taps FIR filters, as mentioned in section 3.2.1, are compared in terms of power and area. All the delay elements are implemented by using arrays of D-type flip-flops. Figure 4.2 shows that when the taps of the filter increased to 128, the segmented architecture with a serial delay line for sub-sequence exhibits power advantage.

The FIR design is split into two parts (i.e. computational and storage part) to get a better understanding of the power distribution inside the FIR design. The detailed power distribution is shown in Figure 4.3. The power reduction comes from the computation part, since the segmented architecture have more balance carry save adder tree.

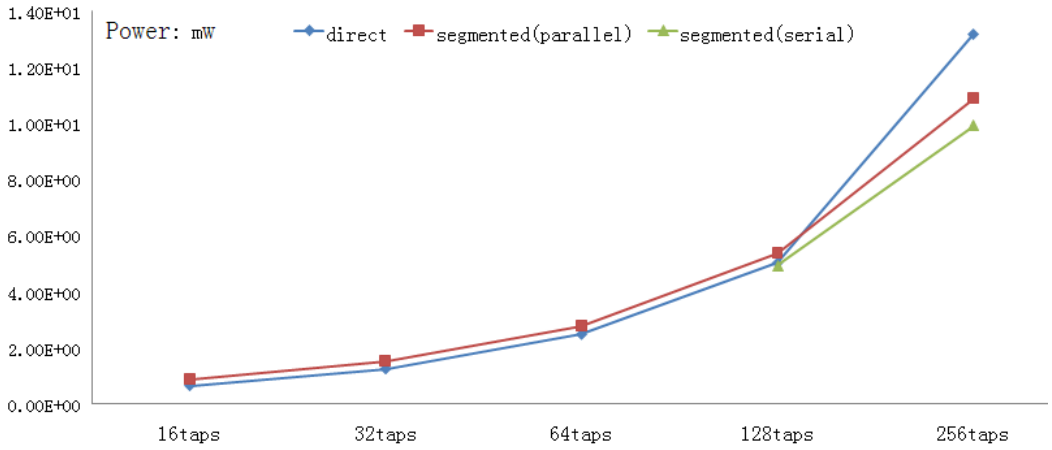


Figure 4.2: Power results on different FIR architectures

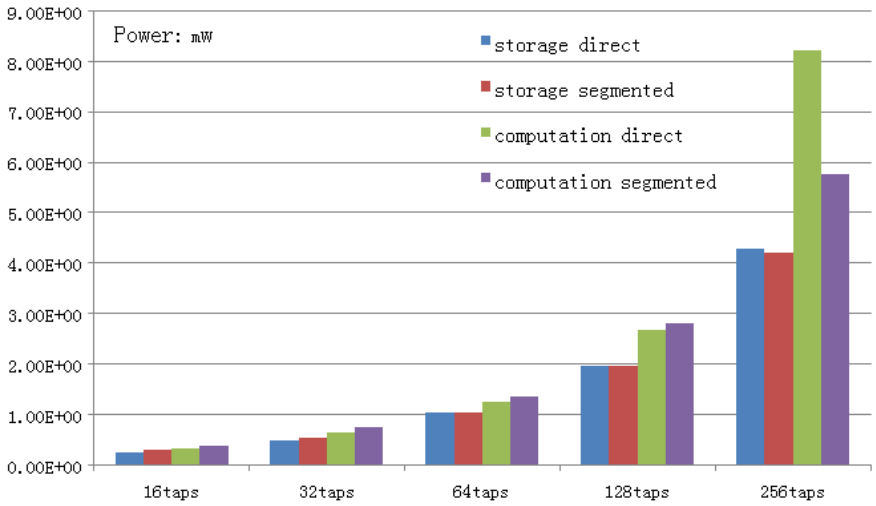


Figure 4.3: Power distribution in direct and segmented FIR architectures

The result of the above comparison serves as the reference for the correlator design implementation. We changed the existing correlator architecture to the one with serial delay elements, as introduced in section 3.3.2. This change decreased the number of needed storage elements significantly. Even though the width of the delay line will increased four bits in this case, however, the depth of register will be significantly reduced and dissipates less power.

Moreover, registers without reset are used in the sub-sequence delay line which can further minimize the area by around 12%. This modification doesn't influence power consumption. The table 4.2 shows the benefits from these modifications.

	size of register (depth*width)	power	area(registers with reset)	area(registers without reset)
Parallel delay line	112w x 32b 224w x 32b 336w x 32b	17.0mw	-	-
Serial delay line	112w x 32b 112w x 36b 112w x 36b	8.68mw	682858 um^2	609694um^2

Table 4.2: Comparison between parallel and serial delay line

4.1.4 Register partitioning

Power improvement on sub-sequence delay line by introducing partitions is shown in this subsection. The graph in Figure 4.4 shows the impact of register partitioning by clock gating and operand isolation, a significant power reduction is achieved for the circular buffer. From the two columns for a three partitioned delay line (second and third column), a power reduction of approximate 62% can be observed. To be specific, clock gating contributes around 10%, while the remaining 52% of power reduction is due to the introduced operand isolation. However, the three portioning method is only effective when the chips per pulse is set to 4 or 16, the reason is mentioned in section 3.3.3. Two equal partitioning cases are explored here, with smaller size in each active part. 25% reduction can be obtained by changing from 4 to 7 partitioning. This results in a power reduction of 67% compared with the original design. If the registers are further partitioned into smaller size, the power benefit could be even more. However, the improvement will become less because more isolation elements (i.e. AND gates) are needed and multiplexer becomes more complex.

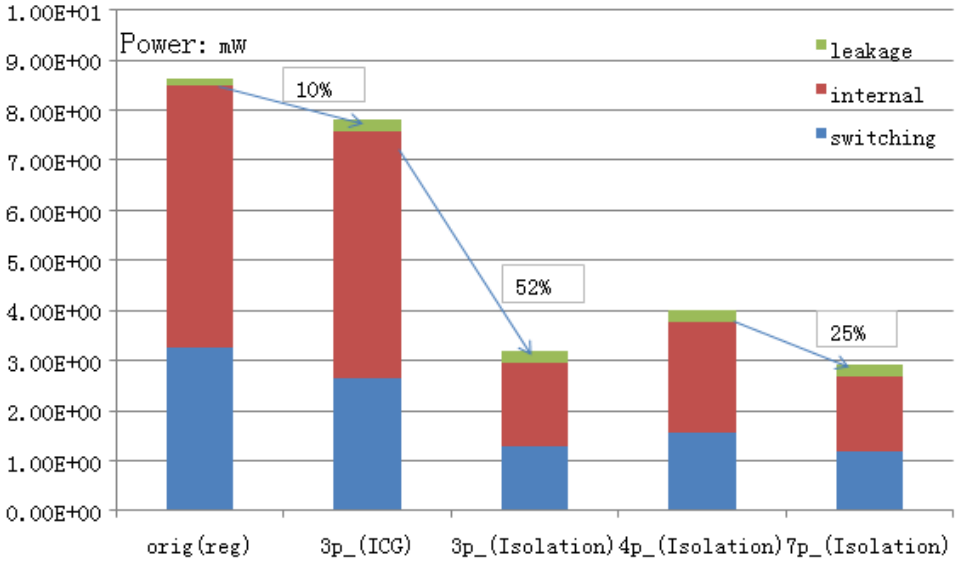


Figure 4.4: Power results on partitioned delay line

As already mentioned in section 3.5, the partitioning of the registers will have a negative

influence on the critical path, since an additional multiplexers will be added between the outputs of registers in the delay line to the following sub-sequence addition. We extract the timing information along the critical path to make a comparison between four partitions and seven partitions. Due to the more partition, the delay between the two-stage multiplexer is increased 0.2 ns after Place and Route. When the standard cells are operating at 840mV supply voltage the standard cells are fast enough to accommodate the additional delay introduced by the additional logic. However, the timing requirements can't be met in case the cells are operating at an even lower supply voltage.

$$Delay_{4p} = T_{clk-to-Q} + T_{1-28mux} + T_{1-4mux} + T_{buffers} = 5.708ns \tag{4.1}$$

$$Delay_{7p} = T_{clk-to-Q} + T_{1-16mux} + T_{1-7mux} + T_{buffers} = 5.921ns \tag{4.2}$$

4.1.5 Biased representation on sub-channels

As shown in Figure 4.5 there is a byproduct power reduction (5%) due to register partitioning. The reason relies on the fact that the way to read the date out of circular buffer has been changed. By changing to biased representation, both from the area as well as power consumption point of view, it gives a modest reduction (10% on average) in multi-operand addition but it has a smaller impact on the whole correlator module. Since the bit switching activities in different sub-channels are different, the power saving isn't equal between the different sub-channels.

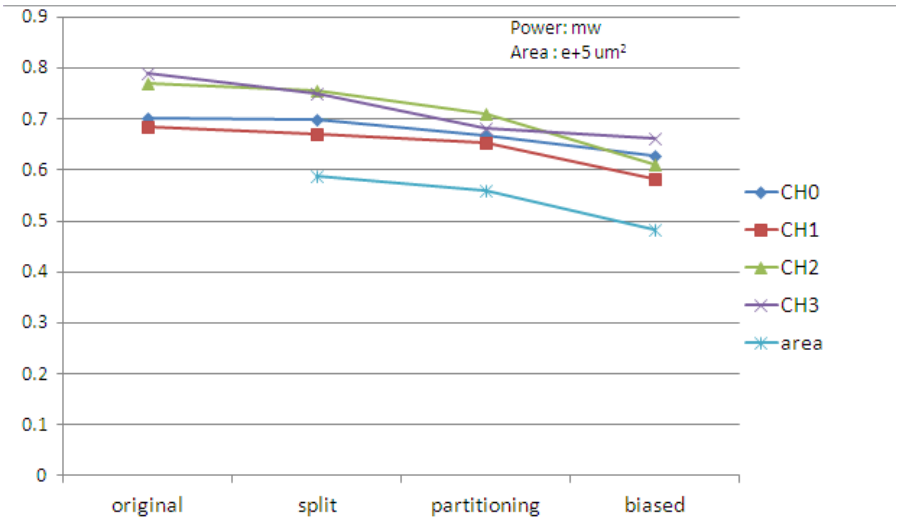


Figure 4.5: Power and area changes of sub-channels in different optimization stages

The sign magnitude calculation is also implemented a comparison, since the data path is duplicated in the addition part but the input width is relatively small in this case, the result of which shows that it is both power and area consuming.

4.1.6 Voltage scaling

According to the equation 2.4, we can obtain a power reduction by 30% due to the supply voltage is scaled from 840 mV to 700 mV for the *subcorr*- and *delcorr* module. The accumulator isn't taken into account since it contains a memory operating at nominal supply voltage.

Since the timing closure is hard to meet after the supply voltage is reduced, pipeline registers are used to alleviate the influence caused by the progressive register partition and shorten the critical path. Intuitively, the introduced registers are regarded as overhead and will increase the total power. However, even at the original supply voltage of 840 mV, the inserted registers can minimize the switching power in the sub-module while the area almost remains the same. As depicted in Figure 4.6, the partial reduction in the pipelined module is 10% due to 20% less switching power and 6% reduced internal power. The expected explanation for this is that the gates along the critical path can have reduced sizes. Also, the additional pipeline stages reduce the glitches and thereby the activities in the design.

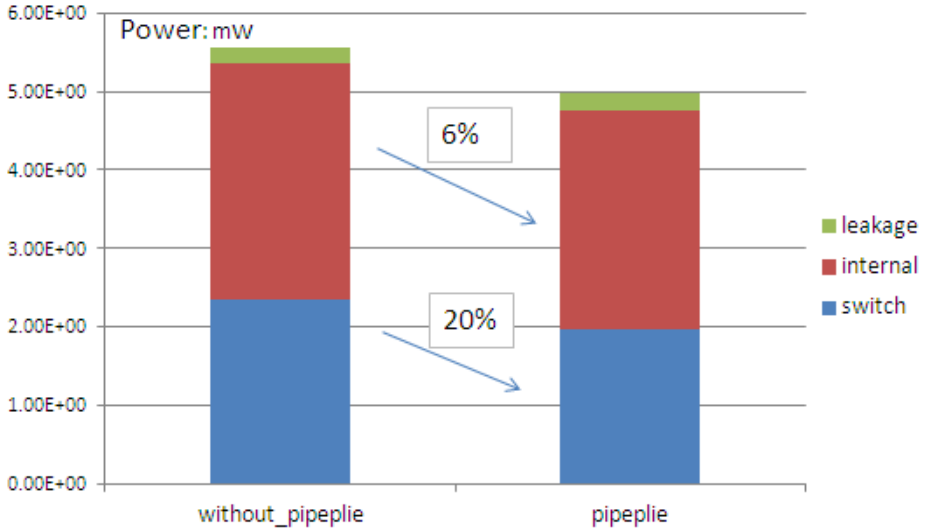


Figure 4.6: Power reduction by inserting pipeline

In order to get the power results under the supply voltage of 700 mV, a new timing library needs to be generated. A standard cell re-characterization tool (i.e. Liberate) provided by Cadence is used to generate the 700 mV library. The total reduction is approximate 3% for full-coherent mode, much less than the expected value, but the area of the total design is increased by 25%.

Figure 4.7 shows that the saving in the *subcorr* and *delcorr* module is limited. If we further zoom in the changes in *subcorr*, we can find that the dissipation in sub-channel (the forth column) is even slightly increased. The reason relies on the fact that the critical path exists in this module now and the gates have increased size to maintain the speed in order to get timing closure.

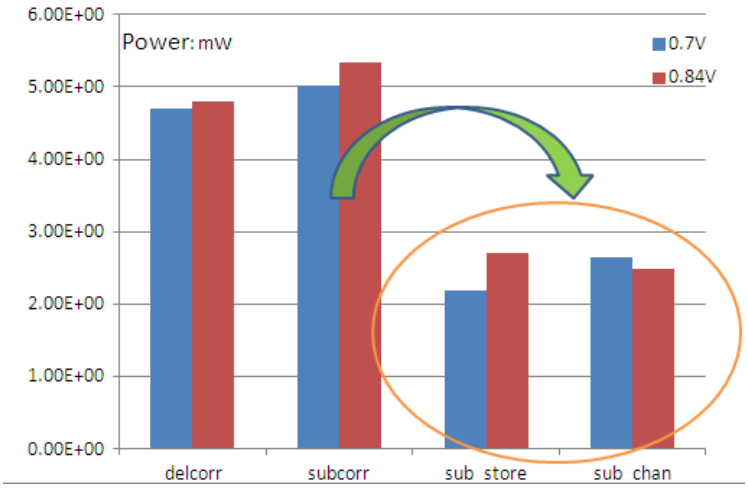


Figure 4.7: Power analysis after the reduced power supply

Figure 4.8 illustrates the normalized area variation in those standard cells based sub-modules. In fact, the clock period of 7 ns is the critical point in the design in terms of timing constraints and power consumption. While the design can be progressively increased to the clock cycle of 5 ns, it is at the penalty of increased logic effort to maintain the performance. To be specific, the increase is moderate during the exploration from 7 ns to 6 ns, but the change turns to be more significant afterwards. It can also be observable that the pipelining introduced in the *delcorr* module can partly mitigate the undesirable effect, so that power benefit still can be obtained.

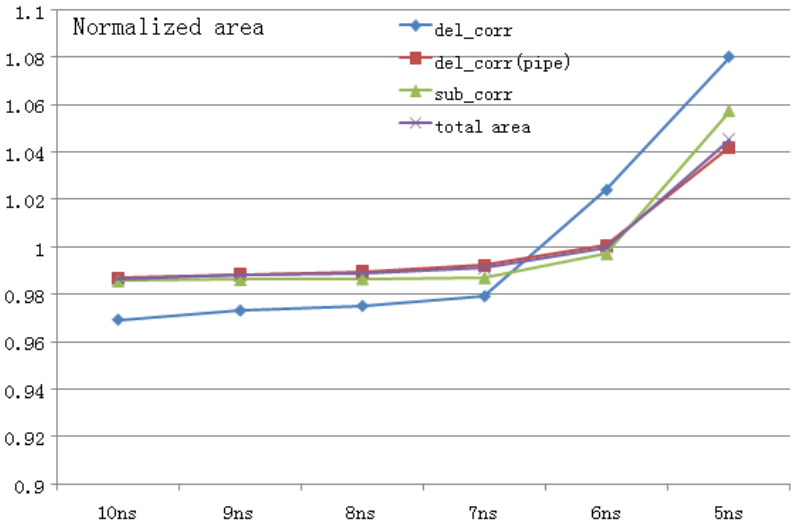


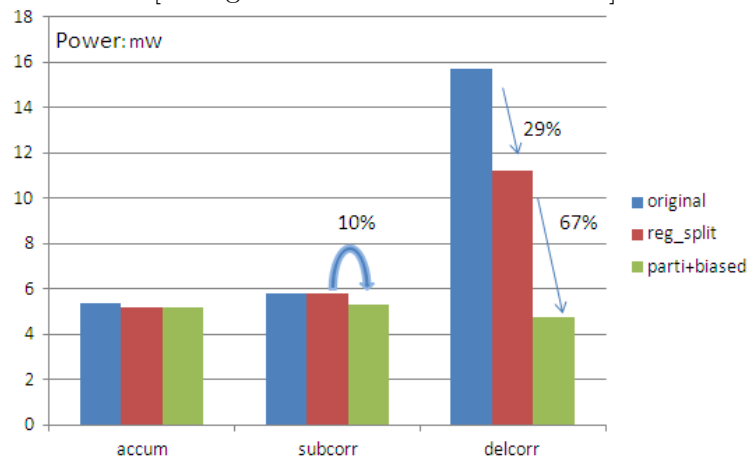
Figure 4.8: Normalized area change during the voltage exploration

4.2 Final power summary

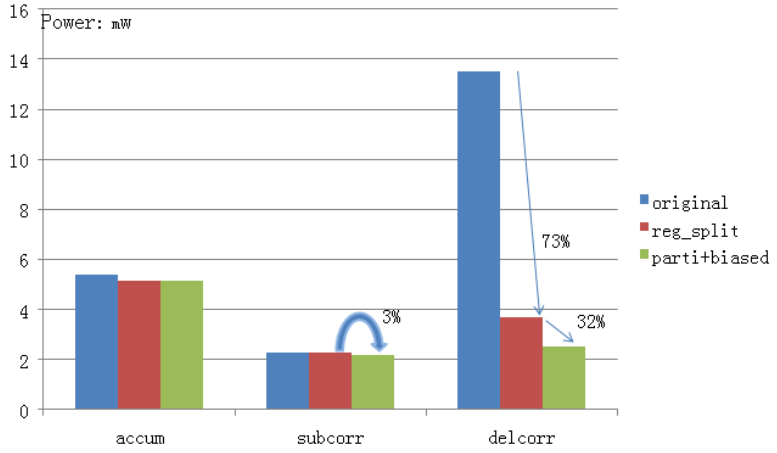
In this subsection, we will give an overview of the power consumption in different modules, and also includes the results in different optimization stages. The results are measured for full-, non- and partially coherent modes, as shown in Figure 4.9(a), (b), (c) respectively. The most significant reduction come from the *delcorr*, especially for the non- and partially coherent case. More than 70% saving can be obtained after the architecture change, since one of the datapaths can be shut down. In case of full-coherent mode, while the benefit of architecture optimization is not as significant as two other cases, two thirds of the reduction is contributed by a register partitioning method. Meanwhile, the configurable value for chip per pulses is also tested between 16, 64 under full coherent case, the power remains constant due to the sub-sequence delay line is equally partitioned. By changing to biased representation, 3% or 10% power dissipation in *subcorr* can be achieved depends on the chosen mode.

While the scope of the work focuses on power reduction on the correlation calculation module without penalizing the performance, the total area and clock tree consumption are also summarized in Table 4.3.

[Subfigure a: Full-coherent mode]



[Subfigure b: Non-coherent mode]



[Subfigure c: Quarter-coherent mode]

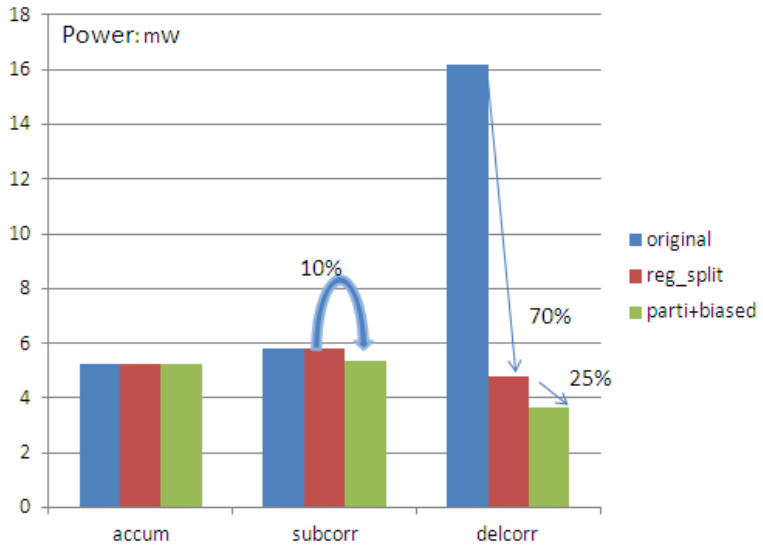


Figure 4.9: Power results in different stages

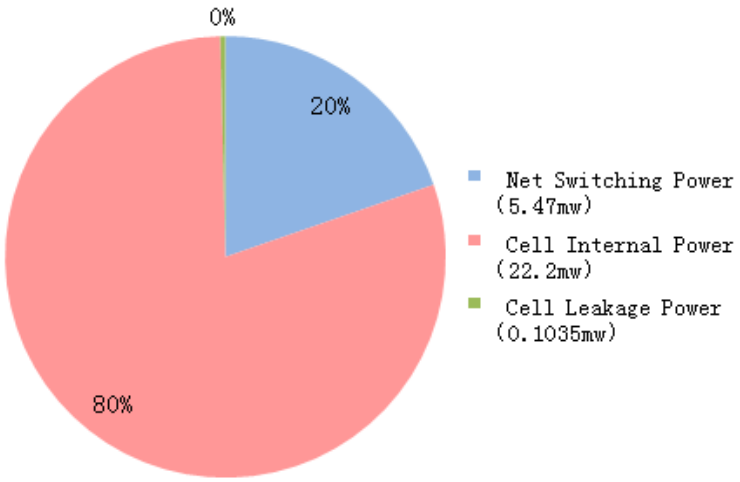
	Original (840mV)	Optimized (840mV)	Optimized (700mV)
Total Power(mw) (full-coherent)	27.8	16.5	15.9
Total Power(mw) (non-coherent)	22.0	10.8	10.6
Total Power(mw) (quarter-coherent)	28.3	15.3	14.8
Area(um^2)	612751	678652	848710
Clock tree	7.0%	11.5%	11.75%

Table 4.3: Final power results summary

It can be seen that the power benefit due to voltage scaling is limited, but the area increase is significant. Besides, it is important to mention that all the power simulations are performed based on the worst-case corner, however, the operating temperature for the standard cell is different between 840 mV and 700 mV version, which are 125 °C and 25 °C respectively. The power result will be even worse if the operating voltage is raised to 125 °C. Therefore, the supply voltage of 700 mV is not a good choice in the design.

The two pie charts in Figure 4.10 shows that the changes in the power distribution between the original design and the optimized one. Due to the mentioned techniques, 10% switching power minimization is achieved and the internal power is halved, but at the penalty of doubled leakage power.

[Subfigure a: original design(840mv)]



[Subfigure b: optimized design(840mv)]

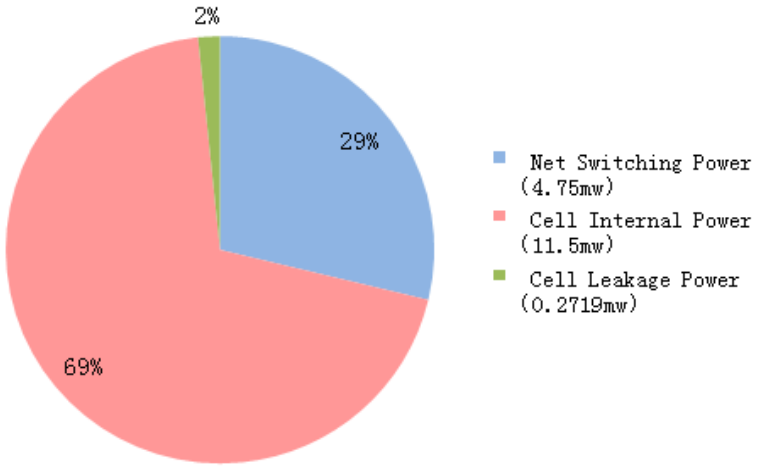


Figure 4.10: Power distribution comparison

4.3 Physical layout view

Figure 4.11 depicts the physical layout view and floorplan layout view after P&R process, which is obtained from SoC Encounter. The dimension of the active area is 1400 x 650 um. The sub-modules are distributed regularly in the schematic. The area highlighted with green is *delcorr* module and the red one is *subcorr* block, accumulator module is located in the left-bottom.



Figure 4.11: Physical Layout of the design

5 Conclusion and Future work

5.1 Conclusion

Within this thesis, a low power exploration of digital correlator is performed for UWB frame synchronization. The work started with a power analysis of an existing preamble synchronization module in the digital baseband and identification on the power distribution, and the result shows that the power dissipation due to memory access donimates the original design. The possibility of architecture optimization is investigated with a comparison between different FIR filters. Based on the result, the data path is further optimized to not only support several modes of operations, but also a varied data flow for different detection modes. The architecture change also provides the possibility to replace the SRAM with D-type flip-flops due to the smaller size of sub-sequence delay line. Register based delay elements are further partitioned by clock gating and operand isolation. Since the storage elements occupy a significant part of the design, the technique by means of register partitioning is pretty useful for overall power. Moreover, the impact of biased representation is studied on the sub-correlators, both area and power for those addition-rich modules can be saved by 10%, but contributes a small reduction for overall design. In order to explore the effect of voltage scaling in the design, pipeling is introduced in the design and results in a tiny power gain because of smaller gate size along the critical path. Theoretically, lower the voltage can effectively bring down the total power. The major handicap with this technique is the penalty of slower cells, which is compensated by changing to the Lvt cells and more pipelining. However, in practice, the improvement is limited to only 3%, the reason for this is that drive strength of the cells becomes larger to meet the timing constraints and consumes more power. From the previous discussions and evaluations of the low power approaches, with maintained power suppy(i.e. 840mV), a possibility for the total power reduction can be achieved up to 2x with 10% overhead on area, the are increase is not significant and can be tolerated for a tape-out.

5.2 Future Work

Regarding the above mentioned evaluations, there are several possible plans to further improvements.

- Further explore the optimum supply voltage. As discussed in the section 4.16, even though the supply voltage can be achieved as low as 700 mV for the standard cell module, it is at the penalty of increased drive strength. Since the Figure 4.8 gives the clue to estimate the optimum voltage, the experiment can be iterated under with a 760 mV voltage supply.

- Fast adder implementation. As shown in Figure 4.7, the power dissipation in sub-channels is even increased by 6% when the supply voltage is reduced to 700 mV. The critical path is now changed to this module, where the multiplication and partial addition are performed between the output from the shared tapped delay line and coefficients. Currently, the carry save adder tree is automatically generated by the CAD tools (i.e. Cadence RTL compiler), and its delay is expected that relatively larger than other adder types, so some faster adders (i.e. carry-lookahead adder) could be tried to keep a balance between timing and power.

- More parallel architecture. There is four-channel complex input are fed into the correlator module in current design, however, as shown in [19], the operating voltage could be even lower to sub-threshold, which is a half of current supply voltage. Hence, we could trade off area and power to further explore the optimum parallelism value.

- Magnitude calculation for biased representation. Biased representation is only applied in the sub-channels for multi-operand addition, while the other modules are based on two's complement data. The reason of this is the existing magnitude module is designed for two's complement format. However, the complex data (e.g. I- and Q) after ADC are inherent biased values and an offset is made to convert them to two's complement. Several conversions can be removed if the magnitude calculation is achieved in the design.

Bibliography

[1] D. Porcino and W. Hirt, "Ultra-wideband radio technology: Potential and challenges ahead," IEEE Communications Magazine, pp. 66-74, 2003.

[2] M. Z. Win, D. Dardari, A. F. Molish et al., "History and Applications of UWB," Proceedings of the IEEE, vol. 97, no. 2, pp. 198-204, Feb. 2009.

[3] B. Allen, M. Dohler, E. E. Okon,W. Q. Malik, A. K. Brown, and D. J. Edwards, Eds., "Ultra-Wideband Antennas and Propagation for Communications, Radar and Imaging." New York: Wiley, pp.1-17, 2006.

[4] G. Durisi and S. Benedetto, "Comparison between Coherent and Non Coherent Receivers for UWB Communications," in EURASIP JASP, vol.2005-3, pp. 359-368, 2005.

[5] J. Zhang, P. V. Orlik, Z. Sahinoglu, A. F. Molisch, and P. Kinney, "UWB systems for wireless sensor networks," Proc. IEEE, vol. 97, no. 2, pp. 313-331, Feb. 2009.

[6] Xiaoyan Wang, Yikun Yu, Benjamin Busze, Hans Pflug, Alex Young, Xiongchuan Huang, Cui Zhou, Mario Konijnenburg, Kathleen Philips, Harmke de Groot, "A meter-range UWB transceiver chipset for around-the-head audio streaming," ISSCC 2012: 450-452, Feb 2012.

[7] Chandrakasan et al., "Low-Power Impulse UWB Architectures and Circuits," Proceedings of the IEEE, vol.97, No.2, pp. 332-352, Feb 2009.

[8] Federal Communications Commission. Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission Systems, Feb 2002.

[9] M. Chiani and A. Giorgetti, "Coexistence between UWB and narrowband wireless communication systems," Proc. IEEE(Special Issue on UWB Technology and Emerging Applications), vol. 97, no. 2, pp.231-254 Feb. 2009.

[10] Alex Young, imec technical notes.

[11] S. R. Aedudodla, S. Vijayakumaran, and T. F. Wong, "Timing acquisition in ultra-wideband communication systems," IEEE Trans. on Vehicular Technology, vol.54, no.5, pp. 1570-1583, Sep. 2005

[12] Neirynck, D.; Philips, K.; Rousseaux, O.; , "Partially coherent signal combination for impulse radio synchronisation," Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, pp.3292-3295, May 2010

[13] David Hodges, Horace Jackson, Resve Saleh, "Analysis and Design of Digital Integrated Circuits," New York, McGraw-Hill, pp.227-237, 2004

[14] A. Fort, J.-W. Weijers, V. Derudder, W. Eberle, and A. Bourdoux, "A performance and complexity comparison of auto-correlation and cross-correlation for OFDM burst synchronization," in ICASSP Proceedings, vol.2, pp341-344, Apirl 2003.

[15] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, Low-power CMOS digital design, IEEE J. Solid-State Circ., pp. 473-484, Apr. 1992.

[16] Architecting, Designing, Implementing, and Verifying Low-Power Digital Integrated Circuits Cadence Design Systems, httpgithub://w2.cadence.com/whitepapers/LowPower_Circuits_WP.pdf

[17] Expanding the Synopsys PrimeTimeSolution with Power Analysis, http://www.synopsys.com/Tools/Implementation/SignOff/CapsuleModule/ptpx_wp.pdf

[18] M. Flury, R. Merz, J.-Y. le Boudec, and J. Zory, "Performance evaluation of an IEEE 802.15.4a physical layer with energy detection and multi-user interference," in Ultra-Wideband, 2007. ICUWB 2007. IEEE International Conference on, pp. 663-668, Sep 2007.

[19] V. Sze and A. P. Chandrakasan. "A 0.4-V UWB baseband processor," Proceedings of the 2007 international symposium on Low power electronics and design," pages 262-267, 2007.

[20] Goto, S., Yamada, T., Takayama, N., Matsushita, Y., Harada, Y., Yasuura, H., "A low-power digital matched filter for spread-spectrum systems," In ISLPED, pp. 301-306, 2002

[21] M. D. Ercegovac and T. Lang, "Low-Power Accumulator(Correlator)," 1995 IEEE Symposium on Low Power Electronics, pp. 30-31, Oct. 1995.

[22] B. Sahoo and K. K. Parhi, "A low power correlator for CDMA wireless systems," J. VLSI Signal Processing, vol. 35, no. 1, pp. 105-112, 2003.

[23] D. Garrett and M. Stan, "Low Power Parallel Spread-Spectrum Correlator," IEEE proc.Circuits Devices Systems, Vol. 146, No. 4, pp. 191-196, Aug 1999.

[24] Jan Rabaey, "Low Power Design Essentials," Springer Publishing Company, Incorporated, pp.317-344, 2009.

[25] Mohit Arora, "The Art of Hardware Architecture: Design Methods and Techniques for Digital Circuits," Springer Publishing Company; pp.91-97, 2012.

[26] Benjamin Büsze, imec technical notes.

[27] E. K. Tsern and T. Meng, "A low-power video-rate pyramid VQ decoder," IEEE J. Solid-State Circuits, vol. 31, pp. 1789-1794, Nov. 1996.

[28] Jairam Sukumar, Madhusudan Rao, Jithendra, Parimala Vishwanath, H. Udayakumar, Jagdish Rao, "Clock gating for power optimization in ASIC design cycle: Theory practice," In International Symposium on Low -Power Electronics and Design (ISLPED), 2008.

[29] S.W. Kim and B. Daneshrad, "A 100 W, 20 Mcps Versatile Correlator Chip for Third Generation WCDMA Systems," in Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers, vol. 1, pp. 130-134, 1999.

[30] K.-Y. Khoo, C.-L. Chen, and A. N. Willson Jr., "Biased two's complement representation for low-power DSP systems," Proc. IEEE 40th Midwest Symp. Circuits Syst., vol. 2, pp.786 -789, 1997.

[31] Harry Veendrick, "Nanometer CMOS ICs: from Basics to ASICs," The Netherlands, Springer, pp.459 -462, May 2008.